RICE UNIVERSITY

A Tale of Two Clades: Unveiling Relationships among *Arabidopsis* and *Lactuca* through Genome Mining in Triterpene Biosynthesis

by

Melisa Moreno Garcia

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Seiichi P.T. Matsuda, Chair
Dean of Graduate and Postdoctoral Studies
E. Dell Butcher Professor
Department of Chemistry
Department of BioSciences

Zachary T. Ball, Associate Professor
Associate Department Chair for
Undergraduate Studies
Department of Chemistry

George N. Bennett
E. Dell Butcher Professor
Department of BioSciences

HOUSTON, TEXAS

May 2016

ABSTRACT

A Tale of Two Clades: Unveiling Relationships among *Arabidopsis* and *Lactuca* through Genome Mining in Triterpene Biosynthesis

By

Melisa Moreno Garcia

The metabolic diversity of natural products in higher plants appears to be driven out of their necessity to adapt and survive in different ecological environments. Among the vast amount of natural products found in nature, triterpenes are widely distributed across higher plants. Triterpenes serve as precursors of essential membrane sterols in primary metabolism and have diverse roles in secondary metabolism, including defense and regulation. These complex chemical structures can be leads in drug discovery.

Oxidosqualene cyclases (OSCs) are enzymes that generate the ring systems of triterpenoids through cyclization of (3*S*)-oxidosqualene. With the arrival of the genomic era, genome mining has become a reliable strategy to give a comprehensive accounting of triterpenoids in organisms whose genome is available. The model plant *Arabidopsis thaliana* has provided through genome mining the first complete picture of triterpene skeletons produced by a single plant species. Now, the availability of different genomes opens up the possibility of exploring the metabolic capabilities of other plant clades. This thesis uses genome mining and heterologous expression in the yeast *Saccharomyces cerevisiae* to explore the biosynthetic diversity of oxidosqualene cyclases in *Arabidopsis* and *Lactuca*, two small clades of the Brasicaceae and Asteraceae families.

OSCs have been widely studied in *A. thaliana* but barely identified in the *Arabidopsis lyrata* genome. Phylogenetic analysis comparing both organisms was used to discover the presence of nine putative OSCs. Three *A. lyrata* OSCs with interesting differences from their *A. thaliana* orthologs were expressed in yeast. The product profile of each enzyme was thoroughly characterized by NMR and GC-MS. *Aly*PEN4 showed a similar product profile to its ortholog in *A. thaliana,* making as its major product thalianol. Analysis of the minor products revealed a novel triterpene, 25-nor-9β-methylpodioda-5(10),17,21-trienol. Expression of *Aly*PEN6 illustrated a notable example of orthologous genes with different product profiles; the ratio of seco-β-amyrin to seco-α-amyrin was

72:0.7 in *Aly*PEN6 and 38:19 in *Ath*PEN4. Finally *Aly*PEN9 proved to be a rather accurate β-amyrin synthase, with no ortholog in *A. thaliana*, which instead performs accurate β-amyrin synthesis with *Ath*LUP4. Interestingly, *Aly*LUP4 is a pseudogene.

Lactuca sativa, the first organism of the Asteraceae family to have its genome available, is a promising model system for studying triterpene synthesis. Here I present progress toward characterizing the triterpene product profile of the *Lactuca* genus in order to understand the evolutionary relationships among different plant clades. Manual splicing of its genome uncovered 17 putative OSCs. An improved product profile of six previously characterized oxidosqualene cyclases from *Lactuca* is reported. Also, a novel cyclase was discovered; it makes nematocyphol as its major product and 39 other compounds, including lupane type rearranged products. One of these compounds, lup-19(21)-en-3β-ol, was found for the first time as part of the product profile of an OSC.

This work shows that *Arabidopsis* and *Lactuca* both have a unique set of oxidosqualene cyclases and make a majority of known OSC products, including β-amyrin, α-amyrin, lupeol, and other common triterpenes. Most of the triterpenes in *Arabidopsis* are also present in *Lactuca*, but the many rearranged lupane skeletons from nematocyphol synthase in *Lactuca* are almost absent in *Arabidopsis*.

Each plant species or clade synthesizes its own set of triterpene secondary metabolites based on selective pressure that guides mutation of existing or duplicated OSC genes to a new OSC that enhances survival and reproductive success. Perhaps OSC genes easily evolve to a new gene for making common triterpenes, whereas OSCs making uncommon triterpenes like nematocyphol evolved through a narrower and less probable set of mutations.

Acknowledgements

My journey through graduate school was filled with many people who helped me become the best I could be, not only as a scientist but also as a human being.

First of all I am deeply thankful to my advisor Professor Seiichi P. T. Matsuda for giving me the opportunity to work with him, and in the process to be able to meet amazing people. Thank you for your guidance and advice, for always being a fair and caring advisor, for your valuable comments regarding my projects, and for your patience in teaching me how to work with sequences and DNA (which I became quite good at). For all of those times that we sat at your office talking about science, for the stories, for all those control experiments that allowed me to became an expert in molecular biology. For believing in me and encouraging my ideas and work around the lab. Thank you for the opportunity that you gave me to become part of your group. I believe that graduate school was so much better because I had the opportunity to work with you.

I would also like to thank Dr. Bill Wilson for all his help and support through my journey. I will miss our wonderful conversations about life and science. Thank you for all your help and input in the development of my projects, for your advice, scientific discussion, for your willingness to help in every aspect of my research, for all the knowledge that you shared with me and for being an invaluable friend.

I owe my deepest gratitude to the Matsuda Lab members, who were not only my lab mates, but my friends. Thank you my Puerto Rican sister Dr. Dorianne Castillo. Your mentorship and work ethic is awesome. I couldn't be happier that I had the opportunity to work with such an amazing person and I am more thankful for your friendship.

Thank you Dr. Paul Bodager for your friendship, it was a privilege to work with you. Thank you for your help with every analytical aspect of my projects, for your input in this manuscript and for been such a good friend.

Jing Jin, it was an amazing day when you decided to join our lab. Thank you for your friendship, for the good conversations about science and beauty and for all your help with my DNA experiments.

Additionally, I owe my gratitude to Matias Kinzurik, Dr. Hui Shan and Dr. Carrie McNeal. Thank you for all the great moments and conversation that we had, for the

knowledge that you kindly shared with me and for all the help and support that I felt when working with you all.

Thank you to all the undergrads that past through the Matsuda Lab, Aparna Bhaduri, Emma Kamaric, Blair Lunceford and Megan Moore. I am also thankful to the people from the Bartel Lab, who one way or another helped me with my *Arabidopsis* and DNA work. Thank you Prof. Bonnie Bartel, Prof. Lucia Strader, Dr. Sarah Burkhart, Dr. Mauro Rinaldi, Kim Gonzalez, Yun-Ting Kao and Pierce Young.

To my thesis committee members Prof. Zachary T. Ball and Prof. George N. Bennett, thank you for agreeing to be part of my thesis committee, for your valuable discussion and interesting ideas.

I am indebted to all my friends who gave me their support throughout my graduate studies and with whom I enjoyed remarkable moments. Thank you Dr. Natalia Gonzales Pech, Dr. Victor Garcia Lopez, Dr. Sergio Dominguez Medina, Dr. Jazmin Godoy Rubio, Dr. Ciceron Ayala Orozco, Lorena Gauthereau and Dr. Andres Rodela. I will always remember all the significant moments that we spent together and I am very grateful for having the opportunity of getting to know you.

I am heartily grateful to my family, for their love and support not only through graduate school but throughout my whole life. They all have supported me unconditionally and encouraged me to fight for what is worth fighting, for whatever it is that makes me a better person, and for those dreams I have attempted to achieve along the course of my life.

Thank you Paulina and Maria Luisa, my very best friends, for all the laughs, the long hour conversations, the advice, the encouragement and the support. I am lucky that God decided to give me such wonderful sisters.

I am especially in debt to my parents Pablo and Maria Luisa, from which I got my best qualities, thank you for your support, love and encouragement.

Finally I would like to thank my very own Dr. Carlos Jimenez Hoyos, you were the one who made me believe that I could aspire to so much more. Thank you for your support, your friendship and your love. You were and are with me every step of the way, through the highs and lows, always encouraging me. I admire you and love you and it was the luckiest day in my life when I met you.

To my husband, Carlos

Table of Contents

List of Figures, Schemes, and Tables

CHAPTER 1

Introduction

1.1 Biosynthesis of Secondary Metabolites

1.1.1 The Role of Natural Products in Plants

Plant secondary metabolites refer to those compounds that are normally not required by a plant to survive and reproduce, but under "special" circumstances play an important role in the adaptation to its surroundings. The environmental challenges differ widely among plants, hence the variety of secondary metabolites found in nature.[1]

Since plants are vulnerable sessile organisms, secondary metabolites are part of their protection; when the plant is threatened by external factors, like pathogens and herbivores, a chemical response is triggered and secondary metabolites are produced.[2]

Secondary metabolites protect plants against microorganisms, herbivores and competing plants.[3] Secondary metabolites can find use as biologically active compounds and drugs.[1] For example, betulinic acid has a cytotoxic effect over cancer cells[4,5] and vaccaroside B is an antimicrobial agent.[6] Secondary metabolites have a lot of possible applications. Their biosynthetic pathways hold a great potential for the production of novel compounds.

Among the wide variety of secondary metabolites that plants can synthesize, one can find terpenoids, a functionally and structurally diverse group in which triterpenoids are included. Triterpenoids are derived from the precursor oxidosqualene (or squalene or dioxidosqualene). As secondary metabolites, triterpenes play a very important role in the survival of plants. For example, the triterpenoid curcubitacine-E that comes from cucumber roots acts as a chemical pesticide against nematodes.[7]

1.1.2 Triterpenes

The word triterpene might be restricted to hydrocarbons, but its common usage includes oxidosqualene cyclization products, such as β-amyrin, the $C_{30}H_{50}O$ alcohol shown in Figure 1.1. Over 100 triterpene skeletons have been identified.[8] Further oxygenated triterpenes and other metabolites are triterpenoids.

Triterpenoids are part of a large and structurally diverse family of natural products called terpenoids, which are derived from C_5 isoprene units. Terpenoids are classified according to the number of isoprene units involved in their construction (Figure 1.1).[9] The biosynthetic pathway for the formation of the subclasses of terpenoids starts with the basic isoprene units found in nature: isopentenyl pyrophosphate (IPP) and its isomer dimethylallyl pyrophosphate (DMAPP), which derive from the mevalonate pathway.[10]

IPP and DMAPP coexist in allylic isomerization equilibrium. These two molecules undergo an electrophilic addition (DMAPP as an allylic cation is attacked by the double bond of IPP) forming geranyl pyrophosphate (GPP, C_{10}). By the same mechanism, subsequent additions of IPP form farnesyl pyrophosphate (FPP, C_{15}) and geranylgeranyl pyrophosphate (GGPP, C_{20}). Head-to-head condensation of two GGPPs makes carotenes, whereas condensation of two FPPs produces squalene (C_{30} triterpene precursor).[9] In eukaryotes squalene is terminally epoxidized to (3*S*)-2,3-oxidosqualene by squalene monooxygenase. Cyclization of oxidosqualene is the key step in the biosynthesis of triterpenes. This reaction is catalyzed by enzymes called oxidosqualene cyclases (OSCs), which fold the substrate and do cyclization via carbocation chemistry.

Classical natural product isolation was originally the foundation for the discovery of novel triterpenes in plants; however the introduction of modern molecular biology techniques allowed the identification and characterization of the enzymes responsible for triterpene biosynthesis. The primary route for the discovery of new triterpenes and OSCs now relies on genome mining approaches[11] and the accessibility to genomic information.[12] *Arabidopsis thaliana*, a plant from the rosids clade, was the first organism from the plant kingdom whose genome was fully sequenced.[13] This model organism became an extremely popular system for studying many aspects of the biology of plants, in particular

characterization of natural products. The next section describes the strategy used in this work to present a complete picture of triterpenes in individual plant species.

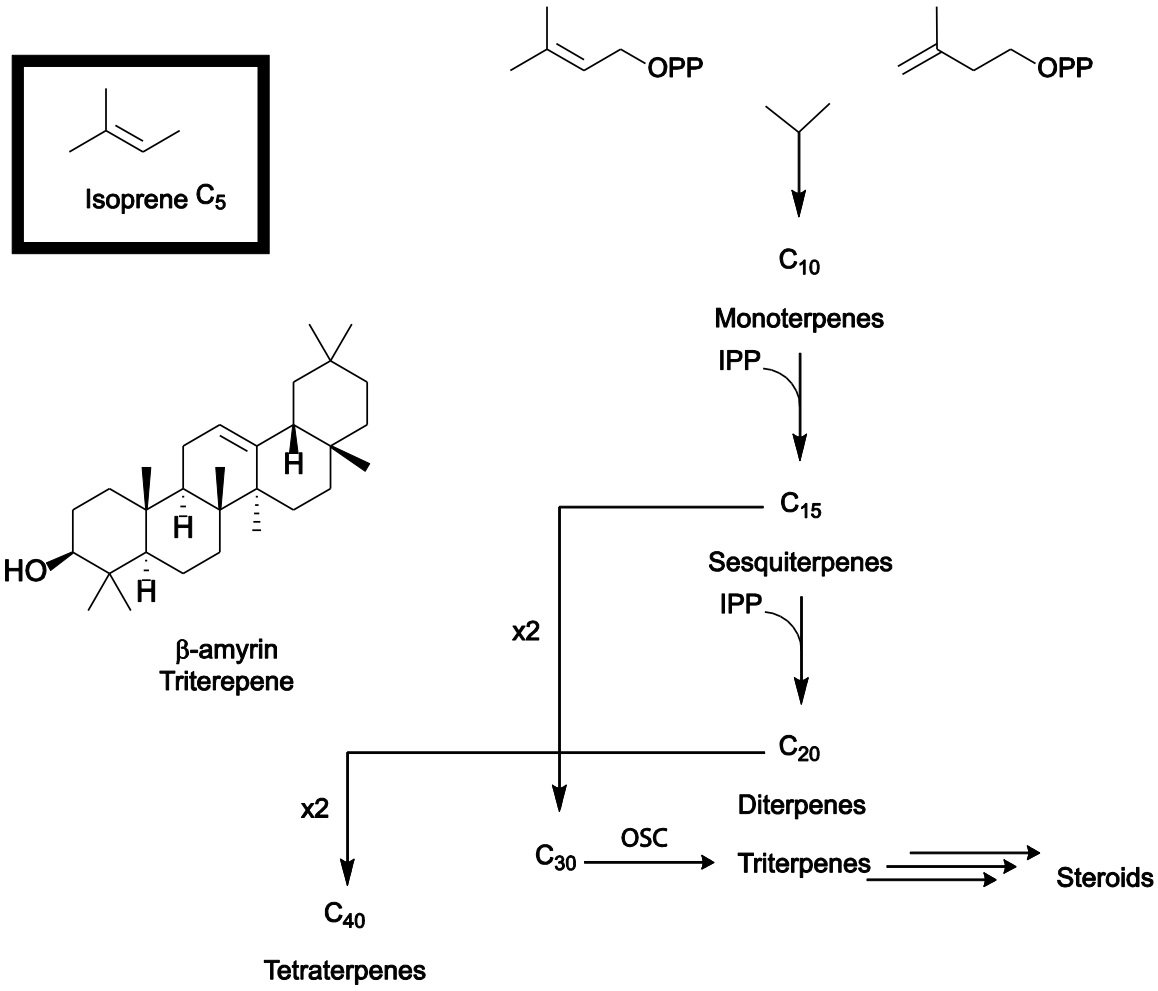


Figure 1.1. Schematic representation of the number of isoprene units involved in the construction of the terpene building blocks.[9a]

1.2 Genome mining in *Arabidopsis thaliana*: Model Organism for the Study of Secondary Metabolism

Arabidopsis thaliana, a model organism in the plant field, has become an extremely popular system for studying many aspects of the biology of plants at the genetic level, including the production of secondary metabolites.[14]

Before genomes of organisms were available, the isolation of new secondary metabolites relied mainly on purification of plant extracts, sometimes guided by the detection of bioactivity. In *A. thaliana* traditional-organic approaches found only four nonsterol triterpenoids: β-amyrin, α-amyrin, lupeol and trinolupeol.[15] This can be explained by the fact that most secondary metabolites are produced only when and where they are needed. Now, with the *A. thaliana* genome available, it is possible to have a detailed picture of the triterpene metabolism in this model organism.

A. thaliana was the first plant whose genome had been fully sequenced. This model system has many advantages for genome analysis, which includes easy cultivation, a rapid life cycle, and a relatively small nuclear genome.[13] Among the enzymes that *A. thaliana* uses to make secondary metabolites are oxidosqualene cyclases (OSCs). Since 1993, OSCs from *A. thaliana* have been studied extensively and have revealed an amazing variety of triterpene skeletons.[16a]

A. thaliana has 13 functional OSCs (Figure 1.2).[11a,16] Over the past two decades elucidating the triterpene product profile of this plant has been an effort pursued by many scientists. The interest over these enzymes is not only the discovery of novel triterpenoid structures, but also the understanding of the enzymatic mechanism by which they are produced. For this purpose, one can use the information revealed by taking a closer look at genomic sequence.

The use of genome mining strategies[11] has allowed the expansion of the number of known triterpenoids that can be found in organisms like *A. thaliana*. This strategy has led to new methodology for studying enzymes and their mechanisms of action, for discovering novel compounds, and for metabolic engineering. Genome mining allows study of genes that are rapidly further metabolized or not routinely expressed. This is a way to find novel natural products that normally are absent or hard to obtain.

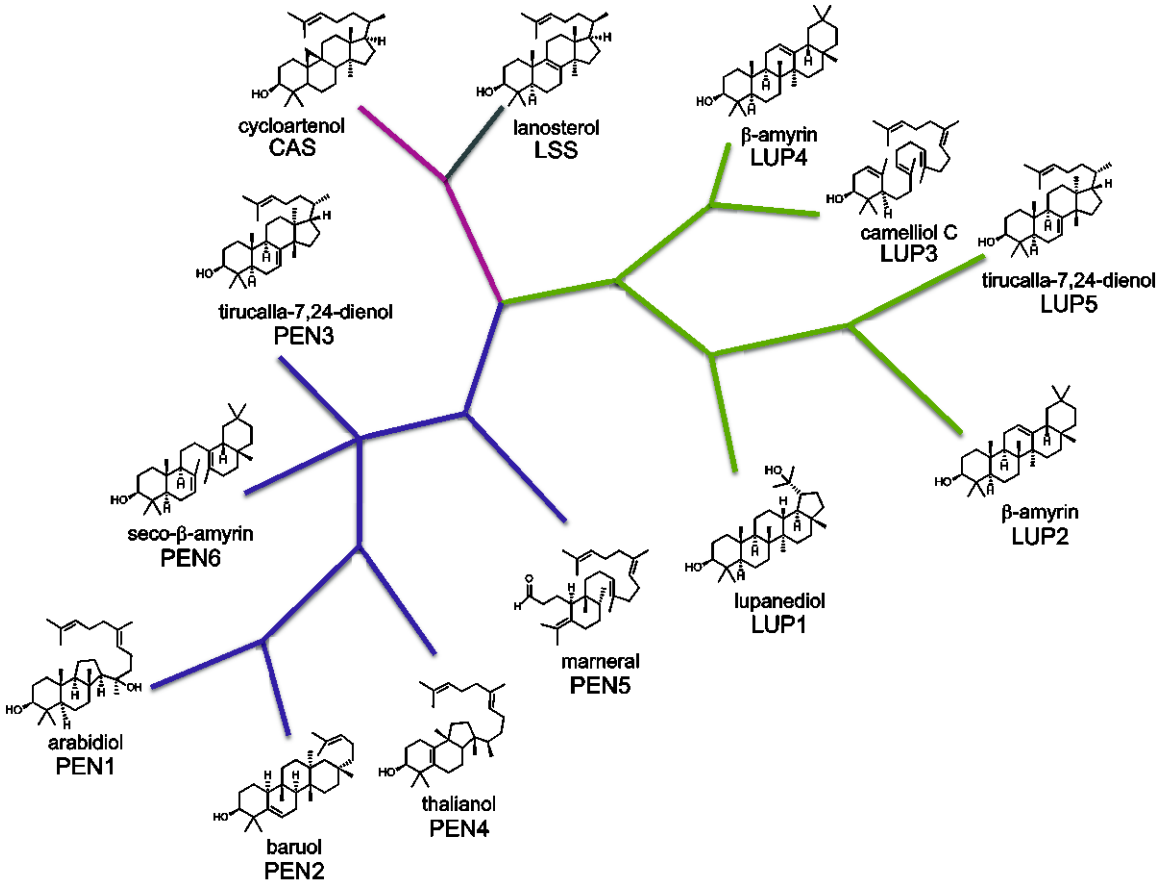


Figure 1.2. The 13 OSCs from *A. thaliana* are separated in three phylogenetic groups CAS (pink), LUP (green) and PEN (purple). Each OSC is represented with its major enzymatic product.

1.3 Diversity of OSCs in Plants

Although Arabidopsis has given a general insight of OSC evolution, the availability of other genomes opens up the possibility for understanding the genetic basis of differences between plants. More than 80 OSCs have been characterized from plants, using heterologous yeast expression and genome analysis.[17]

OSCs play an important role in both primary and secondary metabolism. OSCs involved in the primary metabolism of plants cyclize oxidosqualene via a chair-boat-chair (C-B-C) cationic intermediate to form cycloartenol, the sterol precursor in the plant

kingdom. These enzymes are ubiquitous to the plant kingdom and appear to be the evolutionary ancestor of plant OSCs of secondary metabolism.[18]

Cyclization of oxidosqualene via chair-chair-chair (C-C-C) substrate folding is mainly responsible for formation of triterpene skeletons in secondary metabolism. Some of these OSCs make common triterpenes such as β-amyrin and lupeol, while others generate uncommon triterpenes that have been detected only in a specific family. An uncommon example is shionone, a major product of shionone synthase from *Aster tataricus*.[19] Common skeletons are usually present across families, but each plant species or genus will usually have its own subset of triterpene skeletons.[20] This metabolic diversification in higher plants is often associated with the need to adapt and survive in different environments.[21]

The two plant organisms that are the focus of this work, *A. lyrata* and *L. sativa*, will serve as examples of how often triterpene skeletons between species are conserved and which ones are unique to a given species.

1.4 Comparative Genomic Study: *Arabidopsis* and *Lactuca* Genome

Sequenced genomes offer the opportunity to investigate in depth triterpene biosynthesis. To date, more than 80 genomes from the plant kingdom are available, among them the genomes of *Arabidopsis lyrata*[22] and *Lactuca sativa*.[23]

The first candidate of this comparative study is a close relative to *A. thaliana*. The *A. lyrata* genome, available in 2010, was sequenced with the purpose to expand the understanding of the genus *Arabidopsis*. *A. lyrata* has a relatively small genome (207 Mb) compared with *A. thaliana* (125 Mb), from which it diverged roughly 10 million years ago. My manual splicing of the genome gave 14 putative OSCs. In principle, one would need to study all 14 OSCs to be able to have a comprehensive accounting of the triterpene structures that this plant can make. However, information about triterpene product profile, orthology and expression of OSCs in *A. thaliana* makes the study of this close relative easier.

The Compositae Genome Project (CGP) and the Lettuce Genome Sequencing Project[23] are completing an ongoing effort to sequence the genome of common lettuce,

Lactuca sativa. Related efforts led to the creation of an EST (expressed sequence tag) data set from the *Lactuca* genus, with the expectation that this kind of approach will help to elucidate differences between genomes, and will open the possibility of discovering genes that might confer disease resistance in plants. Once the genome of *L. sativa* was spliced, 17 putative OSCs were discovered, which suggests that triterpene biosynthesis may play a significant role in the secondary metabolism of lettuce. While much remains to be understood about *Lactuca* disease resistance, triterpene metabolites have been known to be involved in resistance to pathogens in other plants.[24]

1.5 Overview

This thesis uses genome mining methodologies and phylogenetic analysis of plant genomes to explore the biosynthesis of triterpenes between two distant eudicot clades.

In Chapter 3, phylogenetic comparison of closely related species illustrates the advantages of model organisms. Identification of novel OSCs in *A. lyrata* was based on their differences to OSCs in *A. thaliana*, without expression of a whole new set of OSCs.

Chapter 4 utilized a different approach to genome mining. EST mining has been used when the genome of an organism is not available. The sequence of nematocyphol synthase was assembled using the basic local alignment search tool (BLAST) to search in *Lactuca* EST libraries for putative OSCs. Creation of EST libraries of the *Lactuca* genus was done with the expectation that, even without a sequenced genome, one can discover genes that might confer disease resistance in plants.

Finally, the theme of Chapter 5 is genome mining of *Lactuca sativa*, a potential model organism of Asteraceae. This chapter presents the progress made towards the characterizing additional OSC product profiles from the *Lactuca* genus.

This work shows that, although *Arabidopsis* and *Lactuca* have a unique set of oxidosqualene cyclases, most of the triterpenes that have been characterized within *Arabidopsis* are also present in *Lactuca*. Contrasting with this convergent evolution, each distant genus contained an OSC making an uncommon set of triterpenes. Additional contrasts include the greater genetic similarity of the studied *Lactuca* species vs. *Arabidopsis* species and the higher loss of functional OSCs in *A. lyrata* vs. *A. thaliana*.

REFERENCES

[1] Wink, M. *Theor. Appl. Genet.* **1988**, *75*, 225-233.

[2] Harrewijn, P.; van Oosten, A. M.; Piron, P. G. M. Natural terpenoids as messengers. A multidisciplinary study of their production, biological functions and practical applications. Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001.

[3] D'Auria, J. C.; Gershenzon, J. J. *Curr. Opin. Plant Biol.* **2005**, *8*, 308–316.

[4] Drag-Zalesinska, M.; Kulbacka, J.; J. Saczko, J.; Wysocka, T.; Zabel, M.; Surowiak, P.; Drag, M. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 4814-4817.

[5] Fulda, S. *Int. J. Mol. Sci.* **2008**, *9*, 1096-1107.

[6] Meesapyodsuk, D.; J, J. B.; Reed, W. D.; Covello, P. S. *Plant Physiol.* **2007**, *143*, 959-969.

[7] Torkey, H. M.; Abou-Yousef, H. M.; Abdel Azeiz, A. Z.; Hoda, E. A. *Aust. J. Basic Appl. Sci.* **2009**, *3*, 4060-4066.

[8] Xu, R.; Fazio, G. C.; Matsuda, S. P. T. *Phytochemistry* **2004**, *65*, 261-291.

[9] (a) Dewk, P. Medicinal Natural Products: A Biosynthetic Approach; Wiley: Chichester, West Sussex, UK, 2002. (b) Jarstfer, M. B.; Zhang, D.-L.; Poulter, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 8834–8845. (c) Pan, J.-J.; Ramamoorthy, G.; Poulter, C. D. *Org. Lett.* **2016**, *18*, 512–515.

[10] Jimenez-Sepulveda, G.; Porta-Duconing, H.; Rocha-Sosa, M. *Rev. Mex. Fitopat*. **2003**, *21*, 355-363.

[11] (a) Fazio, G. C.; Xu, R.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2004**, *126*, 5678-5679. (b) Corre, C.; Challis, G. L. *Chem. Biol*. **2007**, *14*, 7-9. (c) Gross, H.; Stockwell, V. O.; Henkels, M. D.; Nowak-Thompson, B.; Loper, J. E.; Gerwick, W. H. Chem. Biol. **2007**, *14*, 53–63.

[12] Boutamaev, A. M.; Moses, T.; Zi, J.; Nelson, D. R.; Mugford, S. T.; Peters, R. J.; Osbourn, A. *Proc. Natl. Acad. Sci. USA*, **2014**, E81-E88.

[13] The Arabidopsis Genome Iniative. *Nature* **2000**, *408*, 796-815.

[14] Kliebenstein, D. J. *Plant Cell Environ.* **2004**, *27*, 675-684.

[15] Shan, H.; Wilson, W. K.; Phillips, D. R.; Bartel, B. Matsuda, S. P. T. *Org. Lett*. **2008**, *30*, 1897-1900.

[16] (a) Corey, E. J.; Matsuda, S. P. T.; Bartel, B. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 11628-11632. (b) Kolesnikova, M. D.; Xiong, Q.; Lodeiro, S.; Hua, L.; Matsuda, S. P. T. *Arch. Biochem. Biophys.* **2006**, *447*, 87-95. (c) Herrera, J. B. R.; Bartel, B.; Wilson, W. K.; Matsuda, S. P. T. *Phytochemistry* **1998**, *49*, 1905-1911. (d) Kushiro, T.; Shibuya, M.; Masuda, K.; Ebizuka, Y. *Tetrahedron Lett.* **2000**, *41*, 7705-7710. (e) Kolesnikova, M. D.; Wilson, W. K.; Lynch, D. A.; Obermeyer, A. C.; Matsuda, S. P. T. *Org. Lett.* **2007**, *9*, 5223-5226. (f) Shibuya, M.; Katsube, Y.; Otsuka, M.; Zhang, H.; Tansakul, P.; Xiang, T.; Ebizuka, Y. *Plant Physiol Biochem.* **2009**, *47*, 26-30. (g) Ebizuka, Y.; Katsube, Y.; Tsutsumi, T.; Kushiro, T.; Shibuya, M. *Pure Appl. Chem.* **2003**, *75*, 369-374. (h) Xiang, T.; Shibuya, M.; Katsube, Y.; Tsutsumi, T.; Otsuka, M.; Zhang, H.; Masuda, K.; Ebizuka, Y. *Org. Lett.* **2006**, *8*, 2835-2838. (i) Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Kolesnikova, M. D.; Onak, C. S.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2007**, *129*, 11213-11222. (j) Morlacchi, P.; Wilson, W. K.; Xiong, Q.; Bhaduri, A.; Stti-vend, D.; Kolesnikova, M. D.; Matsuda, S. P. T. *Org. Lett.* **2009**, *11*, 2627-2630. (k) Xiong, Q.; Wilson, W. K.; Matsuda, S. P. T. *Angew. Chem. Intl. Ed.* **2006**, *45*, 1285-1288. (l) Shibuya, M.; Xiang, T.; Katsube, Y.; Otsuka, M.; Zhang, H.; Ebizuka, Y. *J. Am. Chem. Soc.* **2007**, *129*, 1450-1455.

[17] Thimmppa, R.; Geisler, K.; Louveau, T.; O'Maille, P.; Osbourn, A. *Annu. Rev. Plant Biol.* **2014**, *56*, 225-257.

[18] Pichersky, E. Gang, D. *Trends Plant Sci*. **2000**, *5*, 439-445.

[19] Sawai, S.; Uchiyama, H.; Mizuno, S.; Aoki, T.; Akashi, T.; Ayabe, S.; Takahashi, T. *FEBS Lett.* **2011**, *585*, 1031-1036.

[20] De Luca, V.; Salim, V.; Atsumi, S. M.; Yu, F. *Science* **2012**, *336*, 1658-1661.

[21] Phillips, D. R.; Rasbery, J. M.; Bartel, B.; Matsuda, S. P. T. *J. Curr. Opin. Plant Biol.* **2006**, *9*, 305-314.

[22] Hu, T. T.; Pattyn, P.; Bakker E. G.; Cao. J.; Cheng J. F.;Clark' R. M.; Fahlgren, N.; Fawcett, J. A.; Grimwood, J.; Gundlach,H.; Haberer, G.; Hollister, J. D.; Ossowski, S.; Ottilar, R. P.; Salamov, A. A.; Schneeberger, K.; Spannagl, M.; Wang, X.; Yang, L.; Nasrallah, M. E.; Bergelson, J.; Carrington, J. C.; Gaut, B. S.; Schmitz, J.; Mayer, K. F. X.; Van de Peer, Y.; Grigoriev, I. V.; Nordborg, M.; Weigel, D.; Guo, Y. *Nat. Genet.* **2011**, *43*, 476-481.

[23] Truco, M. J.; Reyes-Chin-Wo, S; Lavelle, D.; Kozik, A.; Froenicke, L.; Xu, H.; Michelmore, R. *Lettuce Genome Assembly, Structure and Annotation*, Plant & Animal Genome XXII Conference, San Diego, CA, January 11-15, 2014.

[24] (a) Kuang, H.; Ochoa, O. E.; Nevo, E.; Michelmore, R. W. *Plant J.* **2006**, *47*, 38-48. (b) Kuang, H.; Woo, S. S.; Meyers, B. C.; Nevo, E.; Michelmore, R. W. *Plant Cell* **2004**, *16*, 2870-2894. (c) Meyers, B. C.; Shen, K. A.; Rohani, P.; Gaut, B. S.; Michelmore, R. W. *Plant Cell* **1998**, *10*, 1833-1846.

CHAPTER 2

Materials and Methods

2.1 Materials

Restriction enzymes, *BstEII* digest lambda DNA, 100 bp DNA ladder, Gel Loading Dye Blue (6x), Quick Ligation Kit and *Escherichia coli* DH5α (High Efficiency) competent cells were purchased from New England BioLabs (Beverly, MA). pGEM®-T Vector System was purchased from Promega (Madison, WI). PCR was performed with a TripleMaster PCR system (Eppendorf, Westbury, NY) and PCR ExTaq polymerase from TaKara Bio Inc. Gel purifications were performed using the Qiagen Gel Extraction Kit (Qiagen, Inc., Valencia, CA). Bacterial and yeast media components were obtained from United States Biological (Swampscott, MA). Heme (in the form of hemin chloride), ergosterol, bis(trimethylsilyl)trifluoroacetamide (BSTFA), Triton X-100, deoxyribonucleic acid sodium salt from salmon testes (D1626 ssDNA ~2000 bp), Tween 80, and organic solvents were from Sigma-Aldrich (St. Louis, MO). Glass beads for *in vitro* reactions were purchased from BioSpec Products (Bartlesville, OK). Accubond (J&W Scientific/Agilent) and Strata Phenomenex (Torrance, CA) SPE cartridges (1, 2, 5 and 10 g silica) were used with a standard SPE vacuum manifold. TLC silica gel 60 plates (250 µm thickness) were from EMD Chemicals.

2.2 Gas Chromatography-Mass Spectrometry (GC-MS)

Gas chromatography-mass spectrometry analysis of samples was executed using an Agilent 6890N/GC 5973 MSD or a 7890N GC/5975N MSD. The following methods describe the specific conditions for each type of column configuration:

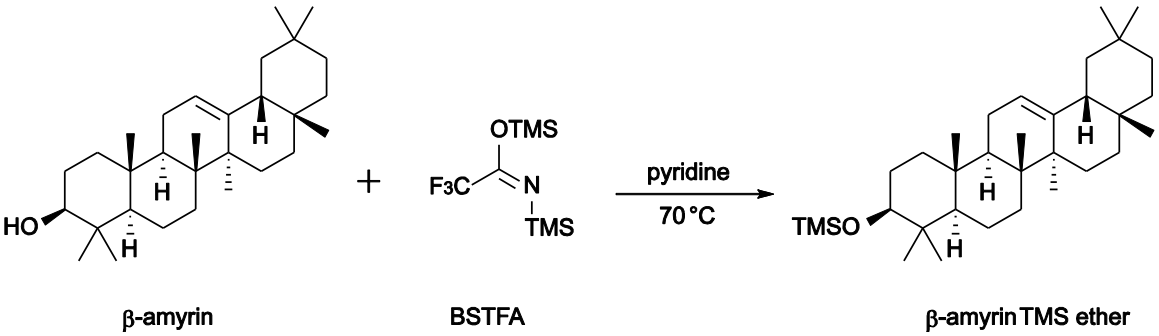
Method A: Agilent 6890N/GC 5973 MSD equipped with a Restek Rxi-35Sil MS 30 m x 0.25 mm x 0.25 µm column, 2.8 mL/min He flow rate, oven at 110 ºC (held 1 min), 30 ºC/min to 270 ºC, 0.5 ºC/min to 280 ºC (held 8 min).

Method B: Agilent 6890N/GC 5973 MSD equipped with a Restek Rxi-5Sil MS, 60 m x 0.18 mm x 0.1 µm column, 1 mL/min He flow rate, oven at 110 ºC (held 1 min), 30 ºC/min to 274 ºC, 0.2 ºC/min to 279 ºC, and 4 ºC/min to 290 ºC (held 3 min).

Method C: Agilent 7890N GC/5975N MSD equipped with a Restek Rxi-5Sil MS, 30 m x 0.25 mm x 0.1 µm column, 2.6 mL/min He flow rate, oven at 110 ºC (held 1 min), 40 ºC/min to 200 ºC, 30 ºC/min to 260 ºC, 1 ºC/min to 275 ºC.

Methods A, B or C (see above) were used to elute the sample components using helium as carrier gas maintained at a constant flow rate. Mass spectra were acquired with electron-impact ionization at 70 eV in full scan mode from 50 to 650 *m*/*z*. Samples (2 µL) were injected splitless mode into an inlet heated to 280 ºC. Triterpene TMS ethers were identified using Agilent Chemstation software by comparing their GC retention time (t_R) and mass spectra against information in our databases and MS libraries.

Trimethylsilyl (TMS) ether derivatives of triterpene alcohols (Scheme 2.1) were prepared by dissolving the sample in diethyl ether, and the desired aliquot of 0.1-10% was transferred to a GC vial containing a 200 µL insert. After evaporation of the solvent, the sample was diluted in 30 µL pyridine, followed by addition of 30 µL BSTFA. The vial was sealed with a cap with Teflon liner, kept at 70 °C for 30 to 45 min and used directly for GC-MS analysis.



Scheme 2.1. TMS Derivatization of triterpene alcohols.

2.3 Nuclear Magnetic Resonance (NMR)

1D (^{1}H, ^{13}C) and 2D (HSQC, HMBC, NOESY and COSYDEC) nuclear magnetic resonance (NMR) spectra were acquired as described in Castillo et al.[1] on 600 and 800 MHz Varian Inova and Bruker Avance spectrometers equipped with cryogenic probes. Samples were prepared with deuterated chloroform ($CDCl_3$) (Cambridge Isotope Laboratories Inc., Tewksbury, MA) that was filtered through activated basic alumina and used to transfer the sample to a thin or thick-walled 5 mm glass tube[2] (Wilmad LabGlass Co., Vineland, NJ). ^{1}H Chemical shifts in 1D and 2D spectra were referenced to internal tetramethylsilane at 0 ppm. ^{13}C shifts were referenced to -0.02 ppm for the 2D experiments or in 1D ^{13}C spectra to the $CDCl_3$ triplet at 77.00 ppm. In HSQC spectra ^{13}C isotope effects on ^{1}H NMR triterpene chemical shifts were minor relative to those of the tetramethylsilane reference and left unadjusted.

For crude extracts of plant material or yeast cultures, sample preparation often required filtration through a silica plug or a 1-g SPE cartridge prior the addition of $CDCl_3$. The crude extract, usually 30 mg in weight, was dissolved in 20 mL hexanes or ether and loaded in a 1 g SPE cartridge, then air or vacuum was applied. The eluent was collected in a 20 mL scintillation vial, evaporated and stored under vacuum to remove any trace of water or solvent.

Preparation of *in vitro* samples for ratio determination often involved small scale saponification. This was often carried out after the removal of oxidosqualene and ergosterol, in order to avoid artifacts that might form during saponification of the crude sample. Since the *in vitro* procedure does not use saponification of yeast cell to obtain the triterpene alcohols, the samples also contained triglycerides and other cell debris that were not completely removed during SPE purification and could interfere with the NMR analysis.

2.4 High Performance Liquid Chromatography (HPLC)

Reversed-phase HPLC separation of complex triterpene mixtures was carried out using an Agilent 1100 HPLC system with a Rheodyne 7125 injector and ultraviolet (UV) detection at 210 nm. C_{18} reversed-phase columns of several sizes were used based on the

amount of material that was injected onto the instrument. For preparative scale, a C_{18} Phenomenex Prodigy column, 250 x 21.2 mm, 5 µm particle size was used. For semi-preparative scale, a C_{18} Phenomenex Sphereclone column, 250 x 10.0 mm, 5 µm particle size, was needed and finally, for analytical scale an Imtakt Cadenza column, 250 x 4.6 mm, 3 µm particle size was used.

Samples were prepared for injection by first dissolving the sample in 200 µL of chloroform ($CHCl_3$), followed by addition of 800 µL of methanol (MeOH) and filtration through glass wool. The final sample (1 mL total) was then loaded onto a 2 mL loop prior to injection. A linear gradient of methanol–water was used to elute the triterpenes out of the column. Flow rates varied from 8 mL/min to 0.5 mL/min based on the column size that was used.

2.5 OD_{600} Measurement of Yeast Cultures with Ultraviolet-Visible Spectroscopy (UV-Vis)

In vitro cultures were harvested using optical density as the measure of saturation. Optical density was measured using a Shimadzu UV-1601 spectrometer at 600 nm. An aliquot of 100 µL was removed from the culture flask under sterile conditions. The sample was placed in a 10 mm UV-grade polymethylmethacrylate (PMMA) cuvette from VWR International Inc. (West Chester, PA.), and diluted with 900 µL of deionized water. Absorbance results were multiplied by a factor of 10 to obtain the optical density of any given culture.

2.6 Centrifugation

Centrifugation (at variable speed) of DNA samples and small scale *in vitro* reactions was done using 1.5 mL microcentrifuge tubes with Eppendorf Centrifuge models 5430R (temperature controlled) and 5415D. Large scale yeast cultures were centrifuged in 50 mL conical tubes using an Eppendorf Centrifuge 5810R, typically at 3800 rpm for 10 min.

2.7 Incubators

Bacterial (*E. coli*) and yeast plates were grown in Fisher Scientific Isotemp incubators at 30 ºC for yeast and 37 ºC for bacterial plates. Liquid cultures of *E. coli* were incubated at 37 ºC in a New Brunswick Scientific Co. G24 Enviromental Incubator Shaker with constant shaking at 250 rpm. For *S. cerevisiae* cultures a New Brunswick G25 or Classic C25 floor shaker was used; both incubators were set at 30 ºC with shaking at 250 rpm.

2.8 Oligonucleotides

Custom designed oligonucleotides for sequencing and PCR amplification were synthesized by Sigma-Aldrich Life Science (The Woodlands, TX). Oligonucleotides were dissolved in Milli-Q water to give a final concentration of 100 pmol/µL. An aliquot of the 100 pmol/µL solution was then used to make a 20 pmol/µL stock solution for PCR reactions. Stock solutions were stored at -20 ºC.

2.9 Polymerase Chain Reaction (PCR)

Polymerase chain reactions (PCR) were executed in an Eppendorf (Hamburg, Germany) Mastercycler Gradient thermocycler. The following set of programs was used for either genomic DNA amplification of a specific exon and/or intron of any given cyclase, DNA mutagenesis, screening of annealing temperature, RT-PCR or addition of suitable restriction sites.

SSLP (A55E030): 39 cycles with initial temperature of 95 ºC, 30 s, annealing temperature of 55 ºC, 30 s, 72 ºC, 30s and final extension of 72 ºC for 5 min.

CAPS (A56E180): 39 cycles with initial temperature of 95 ºC, 30 s, annealing temperature of 56 ºC, 30 s, 72 ºC, 180 s and final extension of 72 ºC for 5 min.

CAPS60 (A60E180): 39 cycles with initial temperature of 95 ºC, 30 s, annealing temperature of 60 ºC, 30 s, 72 ºC, 180 s and final extension of 72 ºC for 5 min.

CAPS60M: 35 cycles with initial temperature of 95 ºC, 30 s, then 94 ºC, 3 min, 94 ºC, 15 s, 50 ºC, 30 s, 72 ºC, 180 s for ten cycles, then 15 cycles of 94 ºC, 15 s, 72 ºC, 180 s + 5 s per cycle and final extension of 72 ºC for 10 min.

CAPSGRAD60: 39 cycles with initial temperature of 95 ºC, 30 s, annealing temperature of 58±8 ºC, 30 s, 72 ºC, 60 s and final extension of 72 ºC for 5 min.

CAPSGRAD120: 39 cycles with initial temperature of 95 ºC, 30 s, annealing temperature of 58±8 ºC, 30 s, 72 ºC, 120 s and final extension of 72 ºC for 5 min.

PCR reactions contained a DNA template (genomic DNA, synthetic gene or cDNA 1-10 ng), a forward primer (20 pmol), a reverse primer (20 pmol), 5 µL of suitable buffer (10 x PC2 or ExTaq buffer) and deoxyribonucleotide triphosphate (dNTP) (10 mmol each). To the PCR mixture ExTaq Polymerase or Taq Polymerase (1-2 U) and 10 x PC2 buffer (5 µL) or ExTaq Buffer (5 µL) were added, followed by Milli-Q water for a total reaction volume of 50 µL.

DNA mutagenesis was generally carried out by introducing the desired mutation into the gene through PCR amplification with modified oligonucleotides. A mutation was usually designed by substituting a codon with minimal change to the DNA sequence. Unique restriction sites used for cloning were found either in the initial DNA sequence within 50 bp of the mutation, or by artificially introducing the site by nucleotide substitution without changing the amino acid sequence, also within 50 bp of the mutation. That the desired mutation was introduced without changing other amino acids was established by sequencing the insert in the expression construct. Products of DNA mutagenesis were ligated to a pGEM®-T Vector System to avoid possible problems with enzymatic digestion.

2.10 Genomic DNA Extraction (Leaf Preps)

Genomic DNA extraction of plant tissue (leaf preps), was done according to Klimyuk *et al.*[3] with minor alterations to the original protocol. One or two leaf pieces were cut and placed into the bottom of a labeled 1.5 mL microcentrifuge tube that was pre-chilled on dry ice. After collecting the plant material on dry ice, it was ground with a pre-chilled plastic microfuge tube pestle, and then returned to dry ice. Once all the leaves (if

multiple samples were prepared) were ground, 10 µL of 0.5 N NaOH was added and left to thaw at RT. The next step, which is critical for the integrity of the sample, was to briefly spin the sample and incubate the samples in a 100 ºC sand bath for exactly 30 s. Once the 30 s are up, 100 µL of LP II Buffer (20 mL of 0.5 M Tris, pH 8.0 with 100 µL 0.5 M EDTA, pH 8.0) was added to neutralize the NaOH in the sample. The tubes containing the genomic DNA extracts were left to rest at – 20 ºC overnight before they were used for PCR analysis.

2.11 RNA Extraction and cDNA Synthesis using Reverse Transcriptase

RNA was extracted from several parts of the plant at different growing stages: seedlings, rosette leaves and flowering plants. RNA extraction was done in RNAse free material and surfaces that were cleaned with 70% ethanol.

4 mL of Tri Reagent per tube (Sigma-Aldrich St. Louis, MO) were added to a 15 mL Falcon tube followed by the addition of selected frozen crushed plant material. The tubes were left to sit at room temperature for ~5 min, then they were centrifuged at 3000 rpm for 15 min and the supernatant was transferred into new 15 mL Falcon tubes. Chloroform was added to each tube (0.4 mL of chloroform per mL of Tri Reagent), and then the closed tubes were manually shaken at room temperature for 5 min.

The chloroform layer was divided into 1-3 microcentrifuge tubes and 0.5 mL of isopropyl alcohol was added per mL of supernatant. The samples were kept at room temperature for 10 min, then they were centrifuged (12 000 rpm, 10 min at 4 ºC) and the supernatant aspirated. The RNA pellet was washed with 1 mL of 75% EtOH and centrifuged again at 7500 rpm for 5 min at 4 ºC. Following centrifugation, the supernatant was discarded and the cell pellet was dried at room temperature for 10 min. The pellet was finally dissolved in 50 µL of DEPC-water.

Prior to the synthesis of cDNA, the RNA template was treated with the Zymo Research DNA-free RNA kit (Zymo Research Corp. Irvine, CA). Once the RNA template was free of DNA, an aliquot (10 pg – 500 ng) of RNA was used with the appropriate primers to synthesize cDNA using the SuperScript III kit (Invitrogen Carlsbad, CA).

2.12 DNA Plasmid Purification

Mini-prep isolation of DNA was done using a modified version of alkaline lysis methodology.[4] An isolated colony from a bacterial transformation plate was inoculated into 2 mL of selective LB media and grown at 37 ºC overnight with constant shaking. The cells were harvested by centrifugation and resuspended in 200 µL P1 buffer (50 mM Tris-HCl pH 8.0, 10 mM ethylenediaminetetraacetic acid (EDTA), and RNase A (0.1 mg/mL)). Once the cell pellet was dissolved, 200 µL of P2 buffer (200 mM NaOH, 1% sodium dodecyl sulfate (SDS) (w/v)) were added to lyse the cells. The suspension was mixed by gently inverting the tube a couple of times and was incubated at room temperature for 5 min. The cell lysis was neutralized by adding 200 µL of P3 buffer (3 M potassium acetate, 1.88 M acetic acid), mixed by inversion and let it sit for 5 to 10 min on ice. The mixture was then transferred into a 1.5 mL microcentrifuge tube and centrifuged. Following centrifugation, the supernatant was transferred into another 1.5 mL microcentrifuge tube. Precipitation of DNA was carried out by adding 1 volume of isopropyl alcohol and incubated at -20 ºC for 30 min or longer to increase DNA yield, after which the plasmid DNA was pelleted by centrifugation for 30 min at 12 000 rpm and 4 ºC. The supernatant was removed by pipet, and the pellet was resuspended in 47 µL of TE8 buffer (10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA). To this mixture, 100 µL of 100% ethanol (EtOH) was added, along with 3 µL of 5 M NaCl. The tube was left on ice for 30 min (or 4 ºC overnight) and centrifuged (12 000 rpm, 10 min, 4 ºC). The ethanol (EtOH) was pipetted off and the pellet was allowed to air dry before it was resuspended in 50 µL TE8. DNA mini preps were stored at -20 ºC.

For preparative DNA purification (maxi prep), a 50 mL *E. coli* culture was grown in selective LB media overnight. Before the cells were harvested, a glycerol was prepared by mixing 0.8 mL of culture with 0.8 mL of 80% glycerol in water solution and was stored at -80 °C. The remaining cells were pelleted by centrifugation, and the DNA was purified from the pellet by ion exchange chromatography using High Purity Plasmid Midiprep System (Marligen Biosciences). The DNA purification was done by following the kit instructions.

2.13 DNA Restriction Digestion

Enzymatic digestion to assess how successful was the plasmid construction was done by taking an aliquot of purified plasmid DNA (1-3 µL). The DNA was then digested in a 25 µL total volume reaction. Preparative restriction digestion reactions for subcloning were performed in 50-100 µL final volume. The digestion reactions details were executed following the manufacturer recommendations (New England Biolabs Catalogs & Technical References).

2.14 DNA Gel Electrophoresis Analysis

Agarose gels for the analysis and purification of DNA fragments were prepared according to the needs of the experiment. Depending on whether a preparative or analytical gel electrophoresis of DNA was being done, agarose concentration of gels changed based on the size of the bands. For 1-8 kbp DNA fragments 1% agarose was used and for fragments yielding bands < 900 bp the agarose concentration was increased to 2%.

TAE buffer was prepared as a 50x stock solution by dissolving 242 g of Tris base, 57.1 g glacial acetic acid, 37.2 g $Na_2{\cdot}EDTA{\cdot}(H_2O)_2$, with pH 8.5 to a final volume of 1 L. Analytical gels were prepared by combining the correct amount of agarose with 500 mL 1x TAE buffer. For preparative gels, 0.283 g of guanosine per 1 L of 1x TAE (GTAE) was added to minimize DNA damage under UV light.

Analysis of DNA samples was done by supplementing each sample with 10x gel loading buffer (0.25% bromophenol blue, 0.25% Xylene Cylanol, and 25% Ficoll 400 in 100 nM EDTA). Samples were loaded into lanes alongside an appropriate DNA marker (100 bp DNA Ladder, 1 kbp DNA Ladder or *Bst*EII digested λ DNA). The run time and voltage used on each gel was adjusted according with the needs of the experiment (ranges from 73 V to 110 V). Progress on the mobility of the DNA bands was monitored by the use of UV light or by the position of the blue dye.

DNA fragments run on preparative gels were excised with razor blades and eluted with QIAquick Gel Extraction Kit (Qiagen) following the manufacturer manual. DNA

fragments run on analytical gels were used for mapping or DNA quantification. *Bst*EII digested λ DNA was used as a reference of size and weight.

2.15 DNA Ligation

Purified DNA bands from preparative electrophoresis gel were ligated using the Quick Ligation Kit from New England Biolabs. When the DNA came from a gel purified PCR amplicon, pGEM®-T Vector System (Promega) was used. The insert to vector ratio was typically 3:1 calculated using NEBioCalculator. However, for small pieces (230 bp - 500 bp) it was recommended to use an 8:1 fold molar excess of insert to vector. The excess of insert to vector minimized the possibility of generating empty vector plasmids during the ligation reaction.

After ligation, the mixture was immediately used for bacterial transformation, except in the case of pGEM®-T Vector System in which the ligation reaction was left overnight at 4 ºC to increase the number of transformants. The expression vectors used for subcloning of all the genes described in this work were pRS426GAL or pRS424GAL.[5]

2.16 DNA Sequencing

Sequencing of all DNA samples used in this work was performed by Lone Star Labs (Houston, TX) or SeqWright, Inc. (Houston, TX). Constructs at various stages of assembly, as well as final expression vectors, were sequenced in order to ensure correct protein expression. Sequencing was also necessary for cDNA and synthesized genes and portions of genomic DNA.

2.17 Bacterial Media

Luria-Bertani [lysogeny] broth (LB media) (10 g/L tryptone, 5 g/L yeast extract, and 5 g/L NaCl) supplemented with an appropriate selection antibiotic was used to grow *E. coli* cultures. The media was sterilized under pressure by autoclaving at 121 ºC for 35 min, and then cooled at room temperature before the addition of the filtered sterilized

antibiotic solution (100 µg/mL ampicillin or 50 µg/mL kanamycin). Solid LB media was prepared following the steps above except for the addition of agar (15 g/L) prior to sterilization. Stock solutions of selection antibiotics (25 mg/mL) were stored at -20 ºC.

2.18 Bacterial Transformation

NEB DH5α Competent *E. coli* (High Efficiency) cells from New England BioLabs were used for bacterial transformation following the protocols provided by the manufacturer. The 5 Minute Transformation Protocol was used for products of ligation because the efficiency is 10% of the High Efficiency Transformation Protocol; this decrease in efficiency allowed us to select individual colonies for Miniprep screening. The High Efficiency Transformation Protocol was used only when the plasmid DNA was verified by sequencing (synthesized genes) or when a different antibiotic (besides ampicillin or kanamycin) was used. When SOC medium was not available, LB media was used. LB agar selective plates were utilized to incubate the antibiotic-resistant transformed cells overnight at 37 ºC.

2.19 Yeast Strains

Two types of *Saccharomyces cerevisiae* yeast strains were used in this work: EHY41[6] and RXY6.[7] *In vivo* production of cyclase products was done using the yeast strain EHY41 (*MAT***a** *upc2-1* pGal1-trHMG1:LEU2). EHY41 is a genetically engineered strain that has the native sterol pathway intact. In addition, it has two important modifications: (1) a dominant mutation in the Upc2p transcription factor (*Upc2-1*) that deregulates the sterol pathway downstream and broadly increases the flux of sterol biosynthesis; (2) an important mutation that allows the overexpression of a truncated version of the hydroxymethylglutaryl coenzyme A reductase (trHMGR1). HMGR performs the rate-limiting step in sterol biosynthesis; overexpression of this enzyme dramatically increases sterol precursor biosynthesis.

Including a foreign OSC in this system means that it must compete with the host's native lanosterol synthase for the substrate (3*S*)-oxidosqualene, but due to the increase in

the carbon flux, the foreign triterpene production should also increase. Expression of OSCs in this system allows for the production and isolation of large amounts of the compounds of interest (10 mg per liter). This is important if one is looking for minor products and we need enough material to perform structure characterization; however, we also get lanosterol metabolites that make analysis and purification a challenge.

In an *in vivo* system a concern is that triterpene products might get further metabolized and thus compromise quantification. Yeast strain RXY6 avoids this and other concerns about *in vivo* systems. RXY6 is a genetically engineered strain that was used for the *in vitro* experiments. It has several modifications that allow expression of our foreign cyclase without exposure to any endogenous substrate. The first modification, a deletion of *ERG7* (lanosterol synthase), removes the native oxidosqualene cyclase. The second modification is a deletion of *ERG1* (squalene epoxidase), thereby preventing epoxidation of squalene to form (*3S*)-oxidosqualene and ensuring that any cyclization occurs after cell lysis, when oxidases are inactivated by low NADPH levels. Those two modifications preclude the formation of the OSC products and their metabolites in vivo. Finally, there is an interruption in the heme biosynthetic pathway, which makes the yeast behave as if it were growing under anaerobic conditions and thus can import ergosterol from the medium. Because the RXY6 strain cannot biosynthesize ergosterol or heme, they must then be included in the medium.

The *in vitro* reaction is carried out by breaking open the yeast cells and introducing racemic oxidosqualene into the cell homogenate at optimal pH. Even though the *in vitro* system provides a clean picture of the product profile of the OSC of interest, it also has some disadvantages. The system is fed with racemic oxidosqualene; most OSCs react almost exclusively with the *S* isomer, leaving the *R* isomer largely untouched. Triterpene artifacts from nonenzymatic cyclization of leftover 3*R*-oxidosqualene are generated during purification and easily confused with authentic OSC products.[8] Thus, oxidosqualene must be removed promptly after incubation. Although, there are no concerns about further metabolism during *in vitro* reactions, there is a clear disadvantage in the amount of recovered since the yield of triterpene alcohols per liter of culture is at most roughly 1 to 2 mg.

2.20 Yeast Media

Preparation of yeast media was customized based on the yeast strain used and the type of plasmid used for expression. YP media (10 g/L yeast extract, 20 g/L peptone) was used for yeast strains that did not carry plasmids. Cultures of yeast strains containing transformed plasmids with selection markers were grown in synthetic complete (SC) media (1.7 g/L yeast nitrogen base, 5 g/L ammonium sulfate, 2 g/L amino acid mixture).

Amino acid stock mixtures were prepared by combining 10 g of leucine with 2 g each of alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, histidine, isoleucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, valine, adenine, and uracil. Amino acid mixtures lacking uracil (SC-Ura) or tryptophan (SC-Trp) were prepared according with the selective marker that was used. Regardless of the media that was used, RXY6 yeast strains had to be supplemented with ergosterol (20 µg/mL) Tween 80 (5 g/L), and hemin chloride (13 µg/mL).

Besides the addition of amino acids, the media also contained either dextrose or galactose as a carbon source with a final concentration of 20 g/L. Solid media was prepared by addition of 7.5 g of agar to the carbon source. Dextrose was used in agar plates or liquid media when cultures were being scaled to the desired volume prior to expression. On the other hand, galactose was used in the last scaling up step as the inducing sugar.

Stock solutions of media components were prepared as 2x solutions, by dissolving components in 500 mL of Milli-Q water and autoclaving at 121 ºC for 20-35 min. Stock solutions of heme and ergosterol were prepared as 100x Heme (65 mg hemin chloride, 25 mL absolute EtOH, 25 mL deionized H_2O and 0.75 mL of 1 M NaOH) and 100x ergosterol solution (20 mg ergosterol, 5 mL absolute EtOH and 5 mL Tween 80). All yeast cultures were grown at 30 ºC, 250 rpm in sterile Erlenmeyer flasks or 50 mL Falcon tubes.

2.21 Yeast Transformation

Yeast transformation of *Saccharomyces cerevisiae* strains EHY41 and RXY6 was done following a modified version of the high efficiency lithium acetate method[9] by Schiestl & Gietz. A 10 mL yeast culture was grown to saturation in YP medium

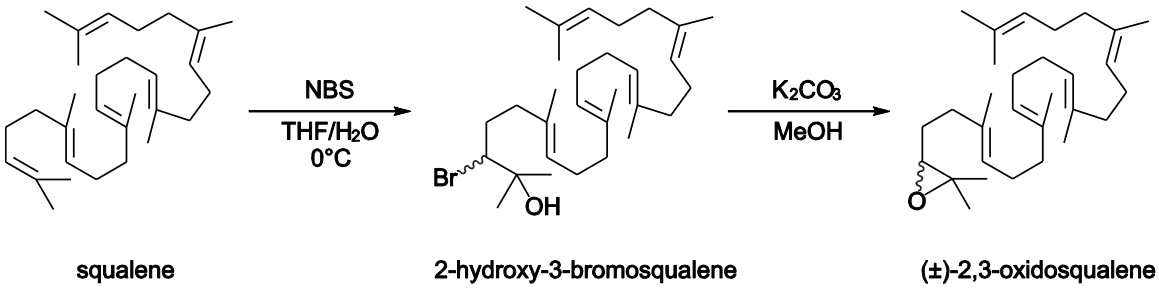
supplemented with heme and ergosterol when required (RXY6). Cells were harvested by centrifugation (3000 rpm, 5 min at 4 ºC). The cell pellet was resuspended in 15 mL of sterile Milli-Q water, centrifuged and the supernatant was discarded. This process was repeated two more times. After the final round of washing, the cells were suspended in 100 µL of sterile water. While the cells were washed, single stranded DNA from salmon sperm (carrier SS DNA) was placed in boiling water for 5 min and chilled in ice while cells harvesting was completed.

Yeast transformation components were added to the 50 mL centrifuge tube containing the newly washed cells in the following order: 100 µL of ssDNA, 25 µL of plasmid DNA (~1 µg of DNA) and 2 mL of yeast transformation buffer (40% polyethylene glycol (PEG) 3350, 0.1 M lithium acetate, 10 mM tris buffer (pH 7.5), 1 mM EDTA, and 100 mM dithiothreitol (DTT). The tube contents were mixed by vortexing and incubated either for 45 min (with vortex mixing every 15 min) at 42 ºC or overnight (12-16 h) at room temperature. Following this incubation 15 mL of sterile Milli-Q water was added, and the cells were pelleted by centrifugation. The pellet was washed two more times. After the supernatant was discarded, the cell pellet was suspended in 1 mL of sterile water and streaked on two selective plates with a sterile pipette tip.

The plates were incubated at 30 ºC until transformant colonies appeared. Yeast colonies grown on the selective transformation plates were then inoculated into 10 mL of selective liquid media with dextrose, grown to saturation and used for preparation of a glycerol stock and for further *in vivo* and *in vitro* assays. Glycerol stock solutions were prepared by mixing 0.8 mL of culture with 0.8 mL of 80:20 glycerol-water and were stored at -80 °C.

2.22 Synthesis of (±)-2,3-oxidosqualene

Racemic oxidosqualene was synthesized from squalene following the procedure described by Nadeau and Hanzlik.[10]



Scheme 2.2. Synthesis of (±)-2,3-oxidosqualene.

2.23 Preparation of (±)-2,3-oxidosqualene Solution

Preparation of a (±)-2,3-oxidosqualene solution was done following the method described by P. G. Bodager.[11] Two hundred mg of synthetic, racemic oxidosqualene were weighed in a tared 100 mL round-bottomed flask, to which 2 g Triton X-100 were added. To this mixture ca. 15 mL methylene chloride were added, and the mixture was mechanically stirred at room temperature until homogeneous. The methylene chloride was then completely removed by rotary evaporation. To this mixture, 7.84 mL deionized water were added to a final volume of 10 mL. This clear solution was mechanically stirred to homogeneity and was stored at -20 °C until use.

2.24 Small Scale *in vivo* Assay

Small scale *in vivo* expression of OSCs was carried out in order to confirm whether a cyclase was successfully expressed. The preliminary results that come from this type of experiment provide insight into how to plan the experiments for quantification and product isolation.

A 10 mL culture was inoculated from a single colony into Synthetic Complete media with 2% dextrose. The culture was grown at 30 °C with constant shaking at 250 rpm. Once it reached saturation, the culture was used to inoculate two 100 mL cultures with Synthetic Complete media and 2% galactose to induce expression. Once the culture reached saturation, the cells were harvested by centrifugation in pre-weighed 50 mL Falcon tubes. After centrifugation, the cell pellet weight was obtained by subtracting the weight of the empty Falcon tube from the final weight (with cells). Once the weight was recorded, the cell pellet was resuspended in 5 mL of 10% KOH (w/v) in 80% EtOH (v/v) per gram of cells, and incubated in a 70 °C water bath for 3 h.

Following saponification, the cell debris was removed by centrifugation, and the supernatant was kept. The volume of the solution was increased by adding 10 mL of water, and the aqueous solution was extracted three times with 15 mL of hexanes. The combined hexanes layers were washed with brine and concentrated by evaporation to dryness in a 250 mL round-bottomed flask. The resulting sample was then transferred into a pre-weighed 20 mL scintillation vial using 10 mL of diethyl ether. The solvent was evaporated under nitrogen, after which the mass of the crude non-saponifiable lipids (NSL) was determined by weight difference and recorded. The sample was dissolved in 1 mL of diethyl ether and a 0.5% aliquot was taken for GC-MS analysis.

2.25 Large Scale *in vivo* Assay

Large scale *in vivo* cultures were used to determine minor products of cyclases, and often provided enough material for the characterization of unknown compounds. The scale of these experiments ranged from 2 L to 8 L and depended of the level of expression of any given cyclase.

Cultures were grown following the small scale *in vivo* procedure up to the 100 mL stage. At that point, the carbon source used in the small scale *in vivo* assay was substituted with dextrose. Once the 100 mL cultures reached saturation, each culture was split and inoculated into 1 L Synthetic Complete media with 2% galactose. Cultures were left to grow under the previously described conditions until they reached saturation.

Once the cultures reached saturation, they were poured into pre-weighed 50 mL Falcon tubes, centrifuged and weighed. The cell pellets were left at -20°C for 1 h, so that it was easier to transfer them into a 1 L autoclaving Pyrex bottle for saponification. The combined cell pellets were resuspended in 5 mL of 10% KOH (w/v) in 80% EtOH (v/v) per gram of cells, covered with aluminum foil and left for 3 h in a 70 °C water bath.

Following saponification, the supernatant was recovered by decanting it into a 1 L round-bottomed flask. The cell debris was then washed once with EtOH and the mixture was left to settle until the cell material was at the bottom of the flask. The EtOH layer was decanted and combined with the rest of the supernatant.

The removal of ethanol is a key step for this procedure and it was done by rotatory evaporation. Ethanol is a significant part of this mixture; large amounts of ethanol present during extractions not only increase the volume of the sample, making it more complicated to handle, but also increase the possibility of forming microemulsions[12] during extraction.

The remaining aqueous phase was extracted with 6 x 50 mL of hexanes. The combined hexanes layers were washed with brine and, dried over anhydrous sodium sulfate, after which the solution was concentrated by rotary evaporation. The mass was recorded and an aliquot from the crude was taken for GC-MS and NMR analysis.

2.26 Small Scale *in vitro* Assay

Small scale *in vitro* assays (100 mL) in the yeast RXY6, were necessary to determine whether the OSC was correctly expressed in yeast and to test the optimal pH for any given enzyme. The pH range tested went from pH 6.0 to pH 7.2.

Initial cultures were grown by inoculating one single colony into 10 mL of Synthetic Complete media with dextrose, supplemented with ergosterol, heme (in the form of hemin chloride) and Tween 80. The cultures were grown to saturation at 30 °C shaking at 250 rpm. After saturation, the 10 mL culture was split in two and inoculated into 100 mL of Synthetic Complete media with 2% galactose as carbon source, and supplemented with heme and ergosterol. These cultures were grown to an optical density (OD_{600}) of ca. 6.0. Once it was reached, the cells were harvested by centrifugation in pre-weighed 50 mL Falcon tubes so that the mass of the cell pellets could be determined. The cells were

resuspended in 1 mL of 0.1 M sodium phosphate buffer per gram of cells. The cells were kept on ice until lysis. A solution of racemic oxidosqualene, 20% Triton X-100 and water was added to the mixture to give a final substrate concentration of 0.5 mg/mL.

The cell mixture was divided in 1.5 mL microcentrifuge tubes (1 mL per tube) containing 200 µL borosilicate glass beads and lysed using a Scientific Industries Genie Cell Disruptor for two periods of 3 min, with one minute resting on ice in between. The enzymatic reaction was allowed to incubate at room temperature for 1 day. An aliquot of the cell homogenate was taken seconds after the procedure was completed and it was analyzed by TLC. The eluents used were first diethyl ether (to separate the triterpenes from the buffer mixture) and then methylene chloride. The TLC plate was submerged in a solution of *p*-anisaldehyde stain and developed on a hot plate.

Three spots corresponding to squalene, ergosterol and oxidosqualene indicated that the substrate had been added. After one day of incubation at room temperature and before the work up of the *in vitro* reaction, a TLC was taken to determine whether it had one or more spots in addition to the initial components; such new spots were expected to correspond to oxidosqualene cyclization products.

The reactions were quenched with two volumes of ethanol, and the cell debris was removed by centrifugation. The aqueous layer was extracted with 3 x 25 mL of hexanes. The hexanes layers were combined in a round-bottomed flask and concentrated by rotatory evaporation. The crude extract was transferred into a pre-weighed scintillation vial with diethyl ether and dried under a stream of nitrogen, after which the mass was obtained. An aliquot was taken for GC-MS analysis, in order to confirm the expression of the cyclase and to get an estimate of the percent conversion of oxidosqualene to enzymatic cyclization products.

2.27 Large Scale *in vitro* Assay

Large scale *in vitro* assays were done to determine product ratios unbiased by metabolism and to compare results with those from EHY41 type experiments; RXY6 was the strain used in these experiments. The scale of the experiments was dependent of the level of expression of each cyclase. The usual scale ranged from 2 L to 4 L.[13]

Initial cultures were grown by inoculating one single colony into 10 mL of Synthetic Complete media with dextrose, supplemented with ergosterol, heme (in the form of hemin chloride) and Tween 80. The cultures were grown to saturation at 30 °C with shaking at 250 rpm. After saturation, the 10 mL culture was split in two and inoculated into two 100 mL liquid cultures. Once the cultures reached saturation they were inoculated into 1 L Synthetic Complete media containing in 2% galactose. Once the cultures reached an optical density (OD_{600}) of ca. 6.0, the cells were harvested by centrifugation in pre-weighed 50 mL Falcon tubes to determine the mass of the cell pellets. The cells were resuspended in 1 mL of 0.1 M sodium phosphate buffer per gram of cells. A solution of racemic oxidosqualene, 20% Triton X-100 and water was added to the mixture to give a final concentration of 0.5 mg/mL substrate. The cells were kept on ice until lysis.

The cell mixture was divided in 1.5 mL microcentrifuge tubes (1 mL per tube) containing 200 µL borosilicate glass beads and lysed using a Scientific Industries Genie Cell Disruptor for two periods of 3 min, with one minute resting on ice in between.

An alternative version of this procedure was used for *Arabidopsis lyrata* OSCs, in which after lysis the contents of the 1.5 mL microcentrifuge tubes were combined into a 50 mL Erlenmeyer flask, and it was then that the racemic oxidosqualene was added. The enzymatic reaction was allowed to incubate at room temperature for 1 day. An aliquot of the cell homogenate was taken seconds after the procedure was completed and analyzed by TLC. The eluents used were first diethyl ether (to separate the triterpenes away from the buffer mixture), and then methylene chloride. The TLC plate was submerged briefly in a solution of *p*-anisaldehyde and heated on a hot plate.

Three spots corresponding to squalene, ergosterol and oxidosqualene indicated that the substrate was added. After one day of incubation at room temperature, and before the work up of the *in vitro* reaction, a TLC was taken to determine whether it had one or more spots in addition to the initial components; these spots will correspond to oxidosqualene cyclization products.

The reactions were quenched with two volumes of ethanol, and the cell debris was removed by centrifugation. The ethanolic supernatants were combined into a round-bottomed flask, and most of the ethanol was removed by evaporation at low heat (to avoid potential decomposition or non-enzymatic cyclization of the excess of oxidosqualene

present in the mixture). The aqueous layer was extracted with 6 x 50 mL of hexanes. The hexanes layers were combined in a round-bottomed flask and concentrated by rotatory evaporation. The crude extract was transferred into a pre-weighed scintillation vial with diethyl ether and dried under a stream of nitrogen, after which the mass was obtained. An aliquot was taken for GC-MS analysis, in order to confirm the expression of the cyclase and to get an estimate for the percent conversion of oxidosqualene to enzymatic cyclization products. Further purification and spectral analysis usually followed after confirmation of successful OSC expression.

2.28 Laboratory Growth Conditions for *Arabidopsis lyrata subs. lyrata*[14]

Arabidopsis lyarata sub. lyrata seeds from ABRC #CS22696 or Hobart, Indiana donated by Greg Livovich and Alyssa A. Nyberg were sterilized and then suspended in 200 μL of sterile 0.1% agar, follow by stratification at 4 °C for seven days.

Under sterile conditions, the seeds (60 per plate) were transferred with a sterile Pasteur pipette into Plant Nutrient Medium (PN) supplemented with 0.5% sucrose plates, taped with one round of micropore tape, labeled and left at ca. 23 °C until germination. On the twentieth day, seedlings were transferred into pots and left in a plant growth room at 23 °C with 24 h light for two weeks, with watering twice a week.

Two week old rosettes were then placed at 4 °C for vernalization[15] in order to help them bolt (grow stem). The plants were kept at 4 °C with 24 h light for four weeks; during this period they were lightly watered every week. After the vernalization process was over, the plants were transferred back to the plant room (23 °C with 24 h light), with watering twice a week. The rosettes started to show stems after one week and flowers after 20 days. During the flowering period, the plants required waterabout three or four times a week to secure good seed production.

Most populations of North America *Arabidopsis lyrata sub. lyrata* are out-crossing/cross pollination plants,[16] which means that to get seeds requires two types of populations, with manual fertilization once the seeds germinate and become adult plants. Additional details and photographs could be found in Appendix B.

2.29 Extraction of Triterpene Alcohols from Plant Tissue

Extraction of triterpene alcohols from plant tissue was done following Shan, H. *et al.*[17] Minor adjustments to the protocol were made according to the needs of the experiment. Fresh aerial tissues from *Arabidopsis lyrata* or *Lactuca sativa* were harvested in order to analyze the triterpene alcohols that had been produced in each part of the plant. For *Arabidopsis lyrata* the following parts were collected: flowers, pedicels, rosebuds, siliques, stem, roots, stem leafs, seeds and rosettes leaves. For *Lactuca sativa* the parts that were analyzed were flowers, stem, roots, leaves and seeds.

The plant tissue was extracted in two steps. First, cuticular wax lipids were obtained by soaking the tissue in hexanes (2 x 20 mL per gram of tissue) for 1-2 h per extraction. This process avoided cutting plant tissue or breaking open any cells. After extraction with hexanes, the plant parts were ground and the internal lipids were extracted with a 1:1 mixture of methanol/methylene chloride (20 mL per gram of tissue). Hexanes extracts and methanol/methylene chloride extracts were kept separate for each part of the plant with the purpose of differentiating the triterpene alcohols present in the circular wax from the ones present inside the cells (internal lipids).

The samples were left overnight with constant magnetic stirring. While hexanes extracts were evaporated to dryness, the methanol/methylene chloride extracts were treated by removing methylene chloride under vacuum, leaving only the methanolic extracts. If saponification was required for these samples, it was done according to the *in vivo* procedure described in this work. Once saponification was done, the samples were allowed to cool to room temperature and the excess MeOH was removed by rotatory evaporation. The remaining aqueous material was then extracted six times with 50 mL of hexanes. The combined hexanes layers were washed with brine and concentrated by rotatory evaporation to dryness in a 250 mL round-bottomed flask. The resulting sample was then transferred to a pre-weighed 20 mL scintillation vial using 10 mL of diethyl ether. Each organic phase was evaporated to a residue comprising the non-saponifiable internal lipids. Aliquots were taken for GC-MS analysis and the samples were purified further when required.

2.30 Purification of Triterpene Alcohols

2.30.1 Solid Phase Extraction (SPE)

Purification of crude extracts from saponification, *in vitro* reactions and plant extracts was done by normal phase SiO_2 SPE chromatography. This step was necessary for further spectral analysis and product isolation. Solid Phase Extraction (SPE) purification is a method that has proved to be reliable for these types of experiments. It is a very reproducible technique that requires less solvent and time than a conventional chromatography column. This type of purification allows separation of any of the crude components that are mixed with the triterpene alcohols. The components of a typical sample are OSC products, linear triterpenes (squalene, oxidosqualene, and dioxidosqualene), native yeast sterols, long-chain lipids, and tocopherols.

Regarding the differences of the accumulated compounds in the yeast strains or plant extracts used for this work, the following strategy was adopted for all of them in order to make the analysis of the fractions faster and to maintain consistency within the method, especially when it was necessary to choose samples for GC-MS analysis.

For large-scale assays (usually ~200 mg crude samples) the size of the SPE cartridges was 10 g for methylene chloride/diethyl ether purification, and 5 g for the diethyl ether/hexanes gradient purification system. The size of the fractions collected during SPE purification was 8 mL.

For the yeast strain RXY6, the main goal was to remove the excess of oxidosqualene from the crude sample, in order to avoid possible non-enzymatic cyclization of oxidosqualene. Previous studies in our laboratory[8] had demonstrated that oxidosqualene cyclizes on silica gel under acidic conditions. On the other hand, purification of crude extracts from the yeast strain EHY41 presents another type of challenge, that is the removal of ergosterol and derivatives that may come from saponification.

In order to begin purification, the crude extracts were loaded into a 10 g SPE cartridge and eluted from the stationary phase using the gradient described in Table 2.1.

Table 2.1. Solvent gradient for a standard methylene chloride/diethyl ether solvent system used with 10 g SiO_2 cartridges.

Fraction	% Et_2O in DCM
1-29	0
30-37	100%

Variation of this method regarding column size and/or number of fraction collected were adjusted according to the necessities of the experiment. The majority of the oxidosqualene (in case of RXY6) and squalene will elute within the first 3 fractions; fractions 4 to 15 will contain triterpene alcohols (and 4,4-dimethyl sterols); fraction 23 to 28 will usually include the majority of ergosterol; most triterpene diols will elute within fractions 30 to 34; and some late diols (like taraxastanediols) will elute in fraction 36. This method removes all the squalene from the sample, as well as a majority of ergosterol. As an added benefit for oxidosqualene cyclases that produce triterpene diols, elution with 100% diethyl ether after 29 fractions of methylene chloride elutes the diol products, providing clean separation of these products in the first chromatographic step.

Subsequent purification of the combined triterpene fractions that came from the methylene chloride gradient is generally performed with hexanes/ether gradients such as the ones presented in Table 2.2. This second SPE serves the purpose of removing the traces of ergosterol, squalene and (when required) oxidosqualene from the sample. Triterpene alcohols usually elute in fractions 18 to 34, fractions 35 to 42 will contain the traces of ergosterol, and triterpene diols elute in fractions 43 to 49; late diols will come in fractions 52 to 55.

Table 2.2. Solvent gradient for a standard hexanes/diethyl ether solvent system used with 5 g SiO_2 cartridges.

Fraction	% Et_2O in Hexanes	Fraction	% Et_2O in Hexanes
1-2	0	27-32	10%
3-6	1%	33-38	20%
7-12	2%	39-46	30%
13-18	3%	47-54	50%
19-26	5%	55-60	100%

2.30.2 Preparative Thin Layer Chromatography (PTLC)

Preparative TLC (PTLC) was used in this work mainly for the recovery of compounds that could not be separated by SPE. Pre-coated Silica Gel 60 (thickness 250 µm) plates were used for all PTLC purification.

The first step of this procedure starts by pre-washing the plates in ether (if the development system is 2:1 hexanes/ether), or methanol/methylene chloride 1:1 mixture (if methylene chloride was used). The wash solvent was left to run up to the top, after which the plate was left to dry overnight at 120 ºC. The next day the plate was taken from the oven and cooled to room temperature. The sample (no more than 20 mg for a 20 x 20 cm plate) was dissolved in a minimal amount of diethyl ether and loaded as a thin line 20 mm from the bottom of the plate, after which the plate was placed in a PTLC chamber with approximately 200 mL of solvent. Once the solvent migrated to a pre-determined distance, the plate was taken out of the chamber and let air dry. A small vertical strip (20 mm) was cut, stained with *p*-anisaldehyde and analyzed for distribution of separated components. The areas of interest were scraped off the glass with a razor blade and collected into small columns, from which the compounds were eluted with diethyl ether. The PTLC fractions were analyzed by GC-MS and NMR.

REFERENCES

[1] Castillo, D. A.; Kolesnikova, M. D.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2013**, *135*, 5885-5894.

[2] Thick-walled tubes were used for samples of up to 5 mg and thin-walled tubes were used for samples 6 to 30 mg with a solvent height of ~35 mm.

[3] Klimyuk, V. I.; Carroll, B. J.; Colwyn, M. T. *Plant J.* **1993**, *3*, 493-494.

[4] *Current Protocols in Molecular Biology*; Ausubel, F. M.; Brent, R.; Kingston, R. E.; Moore, D. D.; Seidman, J. G.; Smith, J. A.; Struhl, K., Eds.; Wiley-Interscience: New York, 1999.

[5] Liu, H.; Krizek, J.; Bretscher, A. *Genetics* **1992**, *132*, 665-673.

[6] Jackson, B. E.; Hart-Wells, E. A.; Matsuda, S. P. T. *Org. Lett.* **2003**, *5*, 1629-1632.

[7] Fazio, G. C.; Xu, R.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2004**, *126*, 5678-5679.

[8] Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Kolesnikova, M. D.; Onak, C.; Matsuda, S.P.T. *J. Am. Chem. Soc.* **2007**, *129*, 11213-11222.

[9] Schiestl, R. H.; Gietz, R. D. *Curr. Genet.* **1989**, *16*, 339-346.

[10] Nadeau, R. G.; Hanzlik, R. P. *Methods Enzymol.* **1968**, *15*, 346-351.

[11] Bodager, P. *Ph.D. Thesis*, Rice University **2014**.

[12] In order to break an emulsion one can collect the emulsion layer in a 500 mL round-bottomed flask and let it sit for at least 2 h at -80 ºC or at -20 ºC overnight.

[13] Some OSCs give higher yields *in vitro* if the OSC is induce at the 100 mL scale. For these cyclases, several cultures were scaled up to 100 mL in synthetic complete media with galactose as a carbon source. They were combined and then, the large scale *in vitro* procedure was followed.

[14] a. The Ohio State University, 2015, Handling *Arabidopsis* plants and seeds. Methods used by the Arabidopsis Biological Resource Center (ABRC), accessed 10 August 2015, <http://www.biosci.ohio-state.edu/~plantbio/Facilities/abrc/handling.htm>b. Germination of most of the seeds occurred after 7 days; however they were transferred to soil after 20 days to allow all the seeds to germinate. c. When watering *A. lyrata* during vernalization, it is important to test with fingers how moist the soil is before adding more water. Over-watering will result in brown/purple leaves and growth of algae. In principle a one-time

watering at the beginning of vernalization (0.5 cm of water left on tray) will be enough for the plants to survive the four week period of vernalization.

[15] Riihimaki, M.; Savolainen, O. *Am. J. Bot.* **2004**, *91*, 1036-1045.

[16] Hoebe, P. N.; Stift, M.; Tedde, A.; Mable, B. K. *Mol. Eco.* **2009**, *18*, 4924-4939.

[17] Shan, H.; Wilson, W. K.; Phillios D. R.; Matsuda, S. P. T. *Org. Lett.* **2008**, *10*, 1897-1900.

CHAPTER 3

Exploring Triterpene Diversity in a Genus of Rosids: Study of *Arabidopsis lyrata* and its Evolutionary Relationship with *Arabidopsis thaliana*

This chapter explores triterpene biosynthesis in the *Arabidopsis* genus, which belongs to the Brassicaceae family of the Rosids clade. Specifically, phylogenetic analysis and OSC product profiles of *A. lyrata* are described.

A. lyrata was studied because its genome has been sequenced[1] and it is closely related to the model plant *A. thaliana*. The high level of knowledge that has been developed over the years in the characterization of *A. thaliana* OSCs simplified the study of *A. lyrata* OCSs.

This project illustrates the use of model organisms and genome mining to elucidate similarities and differences between closely related organisms. What differentiates this study from *A. thaliana* genome mining is that information about orthology between the two species facilitated the characterization of triterpene products in *A. lyrata*, without having to express all of its OSCs.

Phylogeny and active-site amino acid sequence comparisons helped to assess which cyclases will be different enough that their characterization might be important in order to complete the product profile of *A. lyrata*. These key OSCs were chosen for one of two main reasons: because they were not present in the *A. thaliana* genome or because they were different enough from their closest homolog that the product profile might be significantly different.

This chapter includes phylogenetic analysis of the *A. lyrata* OSCs, and the results of yeast heterologous expression of *A. lyrata* PEN4 (thalianol synthase), *A. lyrata* PEN6 (seco-β-amyrin synthase) and *A. lyrata* PEN9 (β-amyrin synthase).

Products are numbered from highest to lowest amount, starting with **1** for the major product of each enzyme. Gene sequences, spectral data, and additional information can be found in the Appendix.

3.1 Background and Previous Work

Before this project became part of my research, Dorianne Castillo, Aparna Bhaduri and Lina Hu tried to characterize the product profile of *Aly*LUP6, *Aly*PEN7 and *Aly*PEN8, three of the OSCs found only in *A. lyrata*. However, none of these OSCs were successfully expressed in yeast. Soon after starting work in this project, I established that the preceding experiments done on *A. lyrata* failed because the synthetic clones ordered were faulty. Sequencing analysis revealed that two synthetic genes had a foreign DNA piece inserted in the middle of the coding sequence, thus making them nonfunctional. First I removed the foreign DNA sections, which were human errors created during the commercial gene synthesis. However, as described later in this chapter, further analysis and experimental results on the repaired sequences indicated that both are pseudogenes. The third commercially synthesized gene was erroneously constructed as a chimera, with an exon from an adjacent OSC. I eventually corrected this problem, but heterologous expression in yeast gave no triterpene products.

After these setbacks, the remaining *A. lyrata* OSCs were extensively analyzed to choose the best candidates for heterologous yeast expression. This analysis involved comparisons of phylogeny and sequence identity with *A. thaliana* OSCs. One criterion for each OSC gene was its overall sequence and active site sequence identity to the closest *A. thaliana* homolog. The second criterion was the *A. lyrata* OSC not having an ortholog in *A. thaliana*. The first criterion was fulfilled by *Aly*PEN4 and *Aly*PEN6 and the second by *Aly*PEN9.

3.2 Experimental Procedures

3.2.1 Phylogenetic Analysis of *Arabidopsis lyrata* Genome

Genomic DNA sequences for *A. lyrata* were downloaded from Phytozome[2] and assembled into 10 different contigs using SeqMan.[3] Manual splicing was carried out by aligning corresponding *A. thaliana* OSC coding sequence with *A. lyrata* genomic DNA in order to find the exons by locating the splicing sites. The *A. thaliana* OSCs were

downloaded from NCBI BioSystems database.[4] Multiple sequence alignments were generated in MegAlign (DNASTAR) from amino acid sequences with Clustal W Method using default parameters. Maximum likelihood trees were constructed using MEGA version 6.[5]

Use of these tools to deduce the coding sequence of *A. lyrata* OSCs raised some concerns. Table 3.1 lists those putative OSCs from *A. lyrata* that had some potential splicing errors that will result in nonfunctional enzymes. All these potential errors were insertion/deletion of nucleotides within the putative coding sequence. The results section of this chapter will address the details of these putative OSCs.

Table 3.1. Putative OSCs with sequence conflict.

Putative OSC	General details of sequence conflict
LUP2	Genomic DNA sequence of putative *Aly*LUP2 from Phytozome has 4 extra nucleotides at the end of the first exon when compare to the *A. thaliana* ortholog. This will add 7 amino acids to that exon, making it longer than the exon in the ortholog in *A. thaliana*.
LUP3	Confirm whether the genomic DNA sequence from Phytozome has an extra adenine at the end of the last exon, this would result in a protein that is 14 amino acids longer than the average OSC length (760 aa).[6]
LUP4	Confirm whether the genomic DNA sequence of putative *Aly*LUP4 from Phytozome does not have a guanosine in the middle of an exon when compare with its *A. thaliana* ortholog; if missing this would generate a frameshift.
LUP5	Confirm if the genomic DNA sequence from Phytozome has two extra nucleotides in the middle of an exon; if so it generates a frameshift.
LUP6	Confirm if the genomic DNA sequence from Phytozome has two extra nucleotides in the middle of an exon; if so it generates a frameshift. Putative *Aly*LUP6 was spliced using as a template its closest *A. thaliana* homolog LUP1.

PEN7	Confirm if the genomic DNA sequence from Phytozome has two extra nucleotides in the middle of an exon; if so it generates a frameshift.
PEN8	Confirm if the genomic DNA sequence in Phytozome does have any errors.

After the experimental data was analyzed, genes highlighted in red appeared to be pseudogenes.

3.2.2 Genomic DNA Experiments

3.2.2.1 Genomic DNA Extraction from *A. lyrata* Leaf and PCR Amplification

Genomic DNA extraction of *Arabidopsis lyrata* was done using a modification of the method "Alkali treatment for rapid preparation of plant material for reliable PCR analysis" from Klimyuk, *et al*.[6] PCR programs are listed in Chapter 2.

3.2.3 Plasmid and Yeast Strain Construction

3.2.3.1 Cloning of *Arabidopsis lyrata* OSCs

3.2.3.1.1 Gene Synthesis of *Aly*PEN4, *Aly*PEN6 and *Aly*PEN9

Table 3.2 shows those genes that were synthesized for heterologous expression in yeast. Criteria for the synthetic genes are described below:

The *A. lyrata* genes were found by doing BLASTN of the closest *A. thaliana* homologs against the *A. lyrata* genome. Our lab revised the computer spliced sequences based on a comparison to *A. thaliana* homologs. Then the gene synthesis company did codon optimization for *Saccharomyces cerevisiae* G/C content, removed direct DNA repeats, and prevented stable RNA secondary structures of the gene. The gene was produced synthetically and put into an ampicillin vector pUC57 by GenScript® or in pMA/pMK vector by GeneOptimizer® with designated restriction sites *SalI* and *NotI*. The plasmids were digested with *Sal I* and *Not I* restriction enzymes, gel purified, and cloned into the yeast expression vector pRS426GAL carrying the *URA3* selective marker. The

resultant plasmids containing the 2.3 kbp ORF are listed in Table 3.2. All the inserts were sequenced to confirm the coding sequence.

Table 3.2. List of *A. lyrata* genes synthesized for this project.

Plasmid	Gene name	Closest homolog	Cloning vector	Company
pMMG13.2	*A. lyrata* PEN4 *Aly*884705/916218	*Ath*PEN4	pUC57	GeneScript
pMMG14.2	*A. lyrata* PEN6 *Aly*477013	*Ath*PEN6	pUC57	GeneScript
pMMG9.1	*A. lyrata* PEN9 *Aly*316749**b**	*Ath*PEN6	pUC57	GeneScript
pMMG6.2	*A. lyrata* LUP6 *Aly*894488/926001	*Ath*LUP1	pMA	GeneArt
pMMG7.0	*A. lyrata* PEN7 *Aly*865365/896876	*Ath*PEN3	pMK	GeneArt
pMMG10.1	*A. lyrata* PEN8 *Aly*316749**a**	*Ath*PEN6	pMA	GeneArt

Sequencing revealed that *A. lyrata* LUP6 and *A. lyrata* PEN7 had 1339 bp of unknown sequence in the middle of the protein coding sequence.

The GeneArt version of *A. lyrata* PEN8 also had problems.The last exon of this synthetic gene belongs to another putative OSC *Aly*PEN9, thus making this construct a chimera. *Aly*PEN8 and *Aly*PEN9 are putative OSCs that are contiguously encoded within the *A.lyrata* genome; *Aly*PEN8 comes first, then *Aly*PEN9. The computer spliced sequence provided by Phytozome[2] was spliced incorrectly keeping most of the *Aly*PEN8 coding sequence, but mistakenly deleting the last exon of AlyPEN8 and most of the exons of *Aly*PEN9 except for the last one. The final construct was a mixture of pieces from *Aly*PEN8 and *Aly*PEN9, thus a chimera. PCR-mutagenesis and synthesis of the correct exon for *Aly*PEN8 was done following methods in Chapter 2. The solution to this splicing problem yielded a new putative OSC *Aly*PEN9.

Table 3.2 shows the name of the genes used for this work in black. Names in orange, are the names used by Phytozome.[2] Phytozome sequences were used often as a reference

for manual splicing and proofreading, but since they were often only fragments of the genes, they are given only as a reference. The final version of each *A. lyrata* gene can be found in Appendix C.

3.2.3.2 Yeast Strains Construction

The plasmids mentioned in the previous section were used to transform *Saccharomyces cerevisiae* strains RXY6 and EHY41 using the lithium acetate method (Chapter 2). RXY6 transformants were selected on synthetic complete medium lacking uracil, solidified with 1.5% agar, and supplemented with 2% glucose, 13 mg/L hemin chloride, 20 mg/L ergosterol, and 5 g/L Tween 80. EHY41 transformants were selected on synthetic complete medium lacking uracil, solidified with 1.5% agar, and supplemented with 2% glucose.

3.2.4 *In vivo* and *In vitro* Experiments towards the Characterization of *Aly*PEN4, *Aly*PEN6 and *Aly*PEN9 Synthases Product Profile

4 L cultures of EHY41 with the respective cyclase (see Table 3.3) were grown in synthetic complete medium lacking tryptophan and containing 2% galactose as a carbon source. The cultures were grown to saturation at 30 ˚C with shaking at 250 rpm. After centrifugation, the cell pellets were saponified with 5 mL of 10% KOH (w/v) in 80% EtOH (v/v) per gram of cells for 4 h at 70 ˚C. Following saponification, the supernatants were separated by decantation, and the excess ethanol was removed by rotatory evaporation. The aqueous layers were then extracted with 8 x 100 mL hexanes. The combined hexanes layers were concentrated to dryness by rotatory evaporation in a round-bottomed flask, then transferred to a pre-weighed scintillation vial and evaporated under nitrogen. The crude NSLs (~170 mg) were analyzed by GC-MS (Figure 3.1) and NMR spectroscopy.

RXY6 cultures (2-3 L) with the respective cyclase (see Table 3.3) were grown in synthetic complete medium lacking uracil and containing 2% galactose as a carbon source. The cultures were grown to saturation at 30 ˚C with shaking at 250 rpm. The cell pellets were harvested by centrifugation when an OD_{600} of approximately 6.0 was reached. The

cell pellets were resuspended in 1 mL of 0.1 M sodium phosphate buffer per gram of cells, and then racemic oxidosqualene (OS) solubilized in Triton X-100 (0.25% final concentration) was added to give a final concentration of 1 mg of racemic OS per mL of homogenate. The homogenates were divided into 50 microcentrifuge tubes (per cyclase) containing ~100 borosilicate glass beads. The cells were lysed by beadbeating using a cell disruptor 3 x 3 min with 5 min intervals on ice. After 24 h incubation at room temperature (ca. 25 °C), the reactions were quenched with two volumes of ethanol and the cell debris was removed by centrifugation. Following centrifugation, the ethanol in the samples was almost completely removed by rotatory evaporation. The remaining aqueous phase was partitioned between hexanes and water and extracted with 6 x 100 mL of hexanes. The hexanes layers were combined, washed with 50 mL of brine, dried over anhydrous Na_2SO_4 to give a crude extract that was analyzed by GC-MS and NMR. In total, three crude extracts were analyzed.

Table 3.3. *In vitro* and *in vivo* experiments of *A.lyrata* OSCs.

	Yeast Strain Constructs	Culture size	pH*	Cell pellet g	Crude mg
A	EHY41[pMMG13.2]	4 L	N/A	75	165
	RXY6[pMMG13.2]	3 L	7	28	170
B	EHY41[pMMG14.2]	4 L	N/A	68	155
	RXY6[pMMG14.2]	3 L	7	29	190
C	EHY41[pMMG9.1]	4 L	N/A	83	177
	RXY6[pMMG9.1]	2 L	7	24	70

* pH of RXY6 incubations; not applicable (N/A) to in vivo yeast cultures (EHY41).

The samples were then purified by SPE (solid phase extraction) chromatography. Each crude was dissolved in 2 mL of hexanes and loaded onto a 5 g SiO_2 SPE column (Phenomenex). Fractions (60 x 8 mL) were collected using diethyl ether (Et_2O)/ hexanes gradients described in Chapter 2.

Fractions containing the triterpene alcohols and triterpene diols (usually but not limited to fractions 19 to 33 for triterpene alcohols and fractions 45 to 50 for triterpene

diols) were combined to yield a sample of triterpene alcohols and one of triterpene diols (Table 3.4). These samples were used for GC-MS, ^{1}H NMR and HSQC analysis.

Combined fractions of RXY6 experiments containing triterpene alcohols were subjected to small scale saponification to eliminate triglycerides and other impurities in order to obtain a cleaner NMR spectrum. This procedure was not necessary for EHY41 experiments, since extractions of triterpene alcohols from this strain involves saponifications as the first step.

Table 3.4. Summary analysis of *A.lyrata* OSCs.

	Yeast Strain Constructs	SPE Fractions triterpene alcohols	Mass mg	SPE Fractions triterpene diols	Mass mg
A	EHY41[pMMG13.1]	22-35	24	40-43	~2
	RXY6[pMMG13.1]	24-35	10	41-43	~1
B	EHY41[pMMG14.1]	21-35	25	43-45	3.5
	RXY6[pMMG14.1]	21-35	10	43-45	~1
C	EHY41[pMMG9.1]	23-30	17.7	34-35	~2
	RXY6[pMMG9.1]	6-8	18.9	10	~2

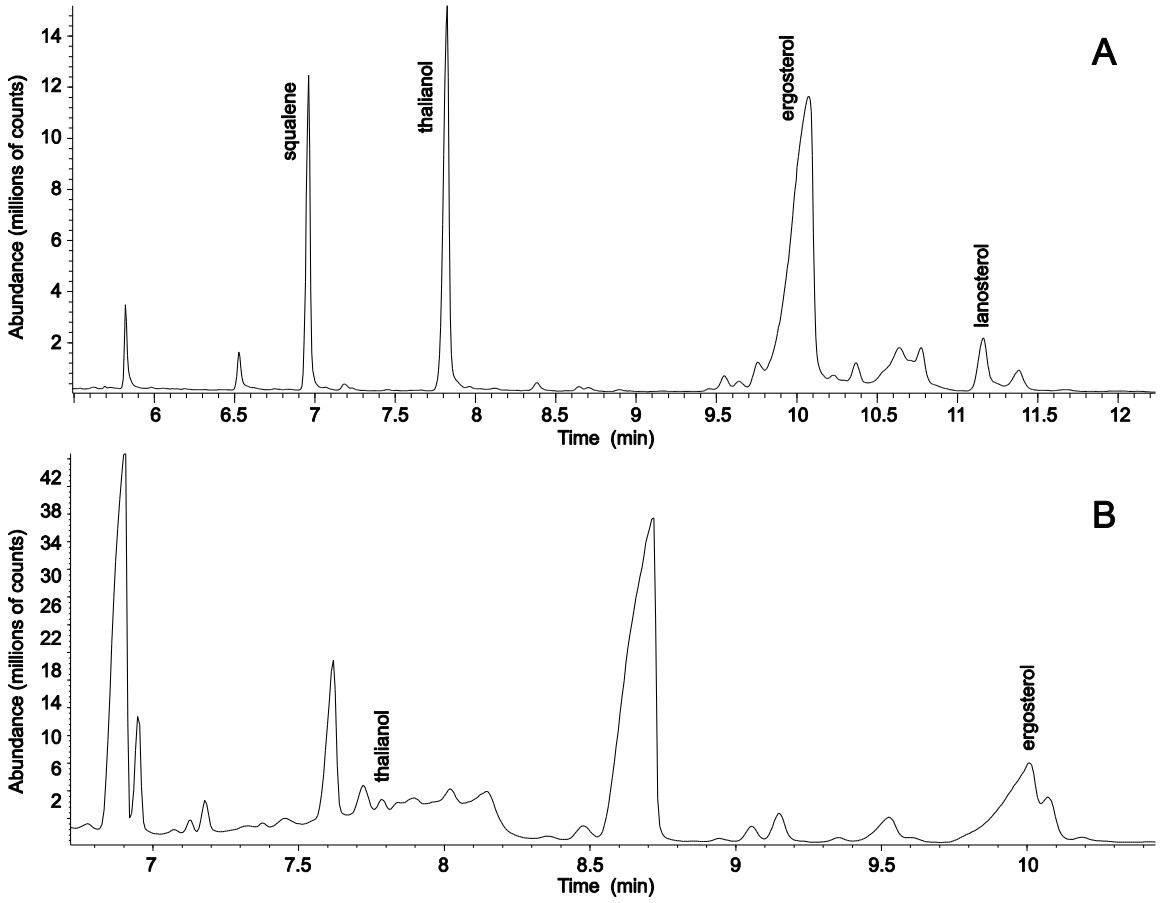


Figure 3.1. A: GC-MS chromatograms of the crude extract of the 4 L EHY41[pMMG13.2] culture. B: GC-MS chromatograms of the crude extract of the 3 L RXY6[pMMG13.2] incubation. pMMG13.2 corresponds to *Aly*PEN4 (thalianol synthase). Retention times in panel B are shorter than in panel A because the GC column was older and thus less retentive.

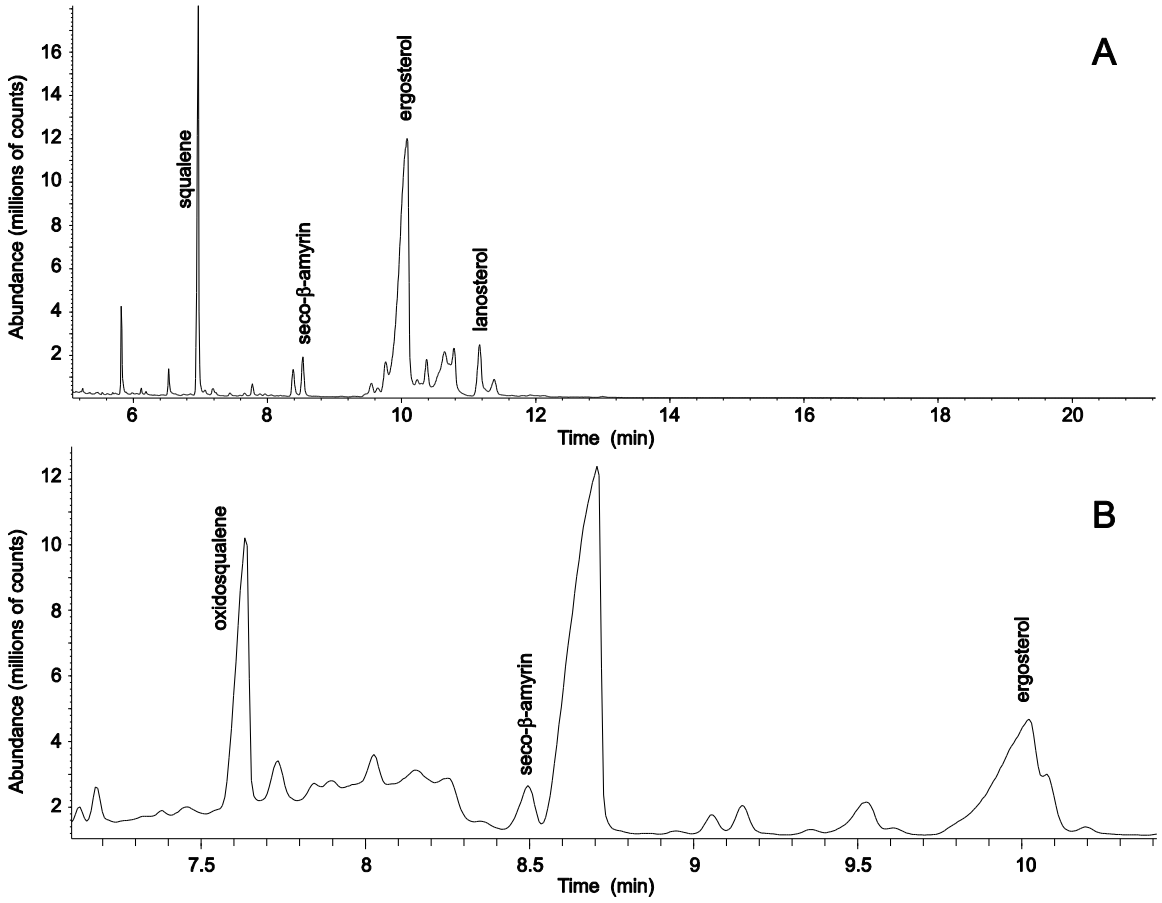


Figure 3.2. A: GC-MS chromatograms of the crude extract of the 4 L EHY41[pMMG14.2] culture. B: GC-MS chromatograms of the crude extract of the 3 L RXY6[pMMG14.2] incubation. pMMG14.2 corresponds to *Aly*PEN6 (seco-amyrin synthase). Retention times in panels A and B due to different ages of the GC column.

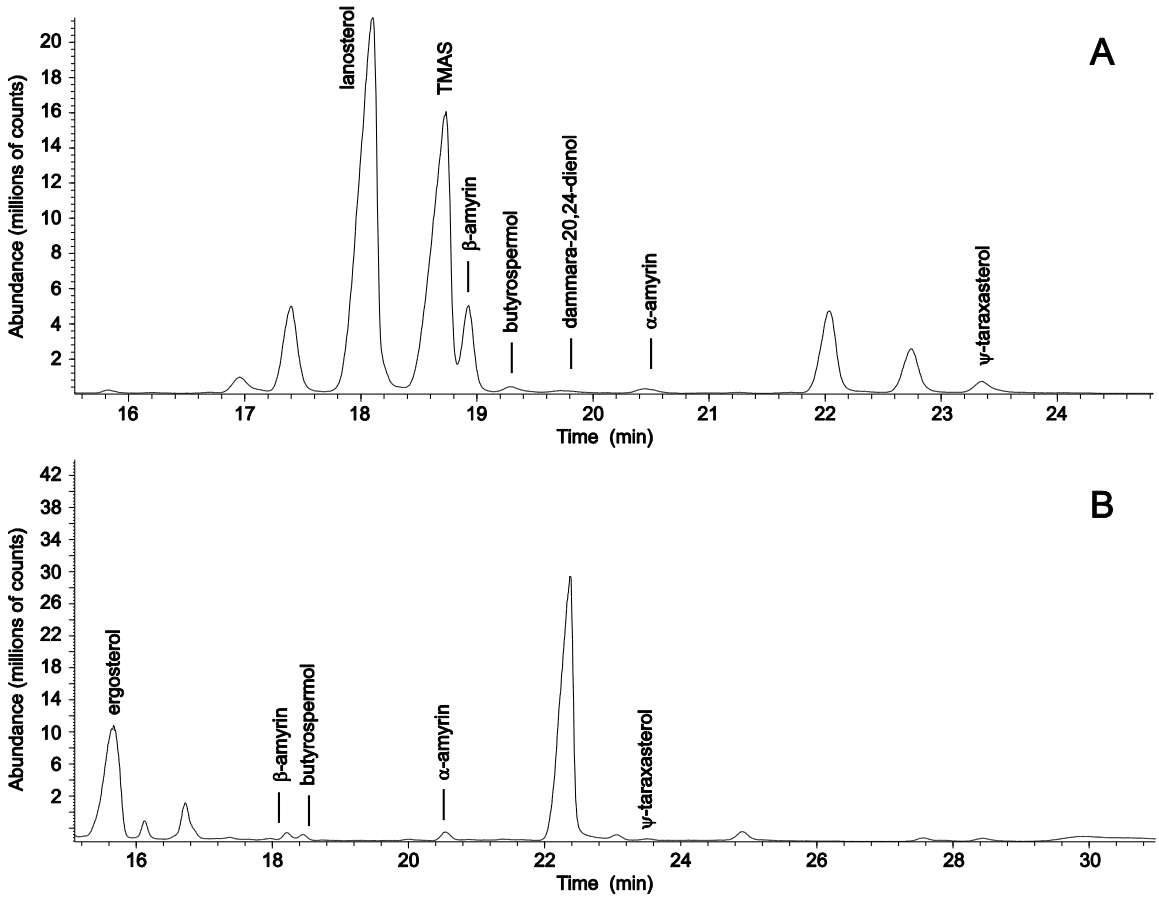


Figure 3.3. A: GC-MS chromatograms of the crude extract of the 4 L EHY41[pMMG9.1] culture. B: GC-MS chromatograms of the crude extract of the 2 L RXY6[pMMG9.1] incubation. pMMG9.1 corresponds to *Aly*PEN9 (β-amyrin synthase).

3.2.5 Extraction of *Arabidopsis lyrata* seeds

A. lyrata seeds (0.5 g) were soaked twice in 20 mL of hexanes. The crude extract was recovered by removing the solvent using rotatory evaporation, leaving behind 8 mg of crude extract. After the hexanes extraction, the seeds were crushed with mortar and pestle and soaked in 10 mL of 1:1 methylene chloride/ methanol per gram of tissue for a day with gentle stirring. Most of the methylene chloride was rotatory evaporated to give a methanolic residue of the modified Folch extract.

After centrifugation, the methanolic extracts were saponified with 10% KOH (w/v) in 80% EtOH (v/v). Following saponification, the supernatant was decanted, and the methanol was removed by rotatory evaporation. The aqueous layer was then extracted with

8 x 50 mL hexanes. The combined hexanes layers were concentrated to dryness by rotatory evaporation in a round-bottomed flask, then transferred to pre-weighed scintillation vials and evaporated under nitrogen. The non-saponifiable lipids (NSLs) crudes were weighed and analyzed by GC-MS and NMR spectroscopy. Further purification was done following the procedures outlined in Chapter 2.

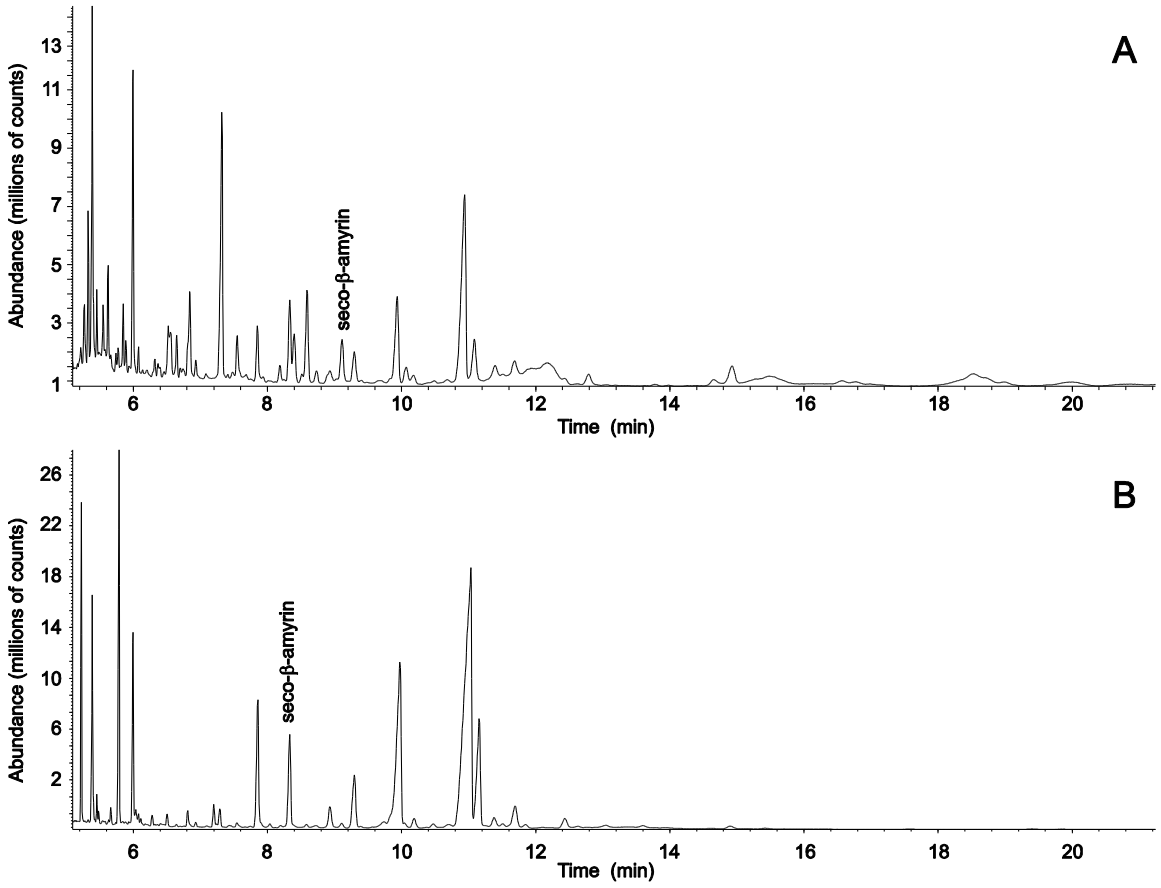


Figure 3.4. GC-MS chromatograms of the crude seed extracts from *A. lyrata.* A: the crude hexanes extract of seeds surface. B: the crude NSLs of the Folch extract.

3.3 Results

3.3.1 Phylogenetic analysis of OSCs from *A. lyrata*

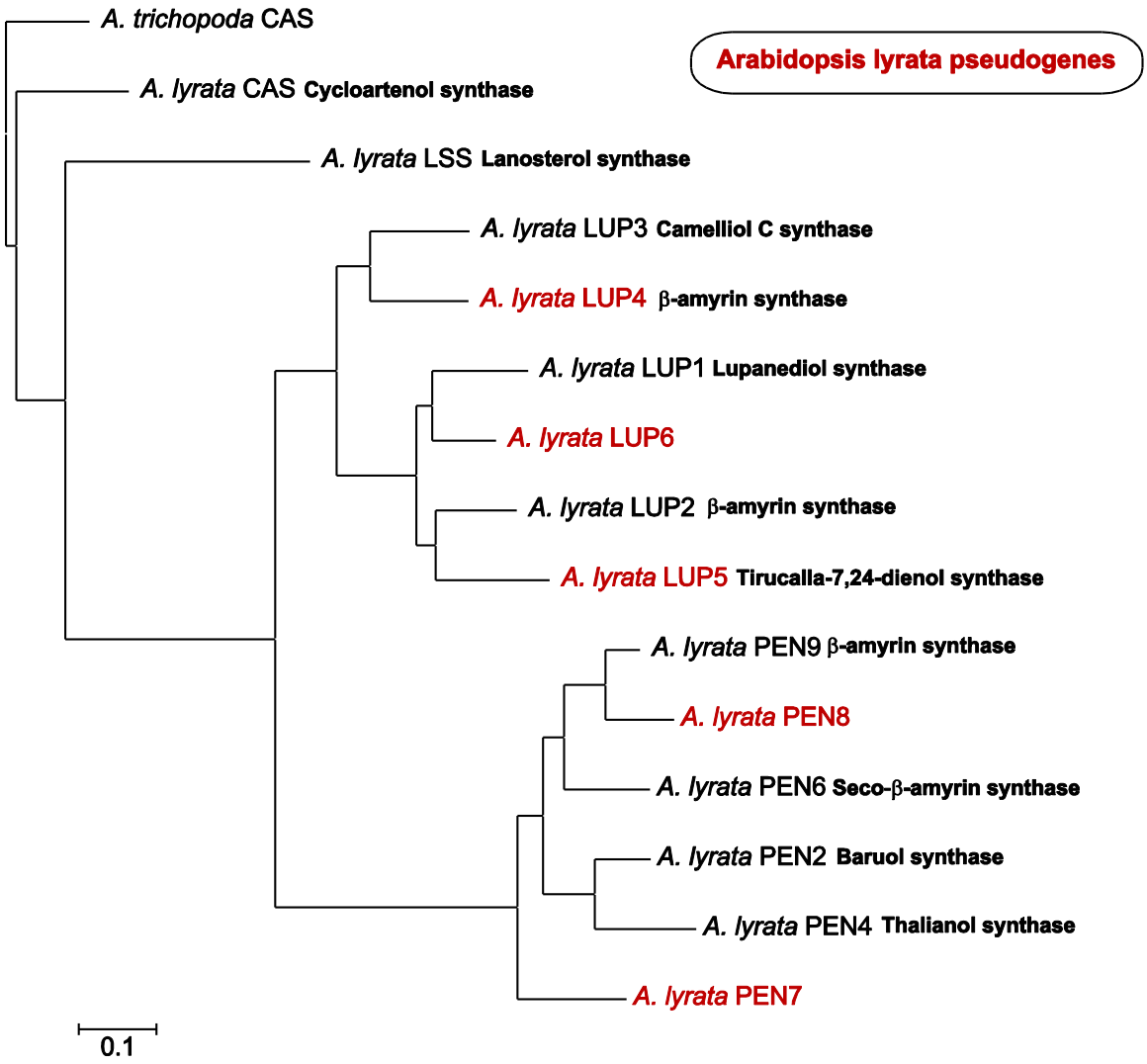


Figure 3.5. Phylogenetic tree of *A. lyrata* OSCs derived from the initial manual splicing of the genome. The tree was constructed by maximum likelihood using MEGA version 6.[5] Multiple sequence alignments were generated in MegAlign from amino acid sequences with Clustal W Method using default parameters. Based on experimental evidence and/or sequence analysis, genes in red appeared to be pseudogenes.

Analysis of genomic sequences and computer-spliced cyclases from *A. lyrata* indicated nine putative OSCs. The Experimental and Discussion sections of this Chapter explain how it was decided which genes were considered full length functional OSCs. Figure 3.5 shows the phylogenetic tree of *A. lyrata*, and Figure 3.6 gives the active site sequence alignment for the various cyclases. Similar phylogenetic analysis and sequence alignments comparing *A. lyrata* with *A. thaliana* are presented in the Discussion section.

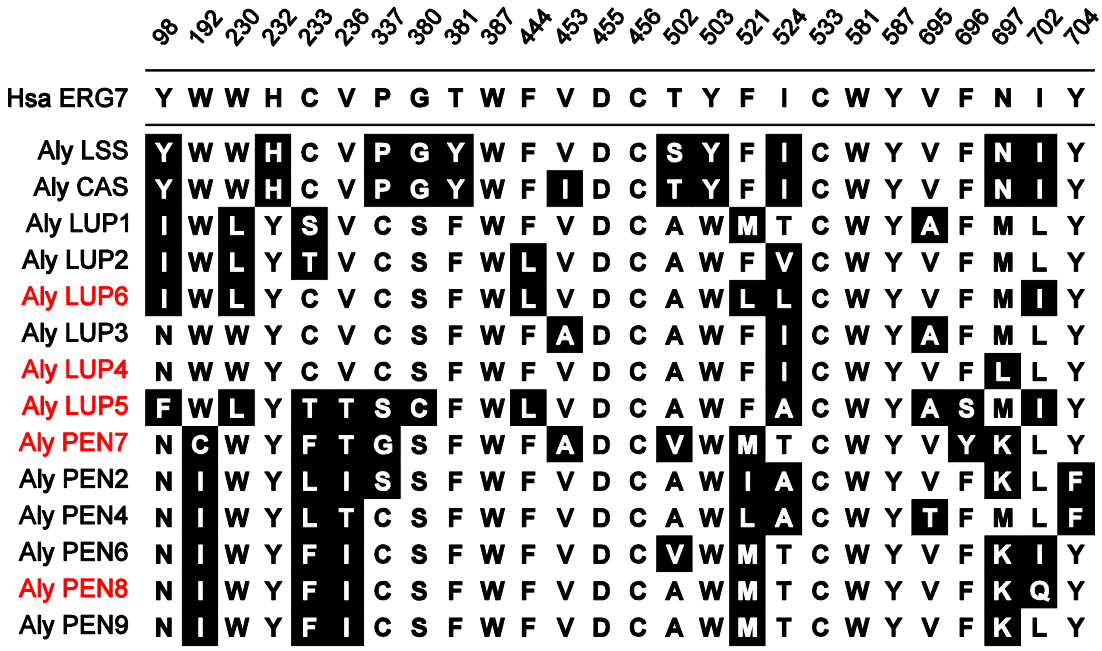
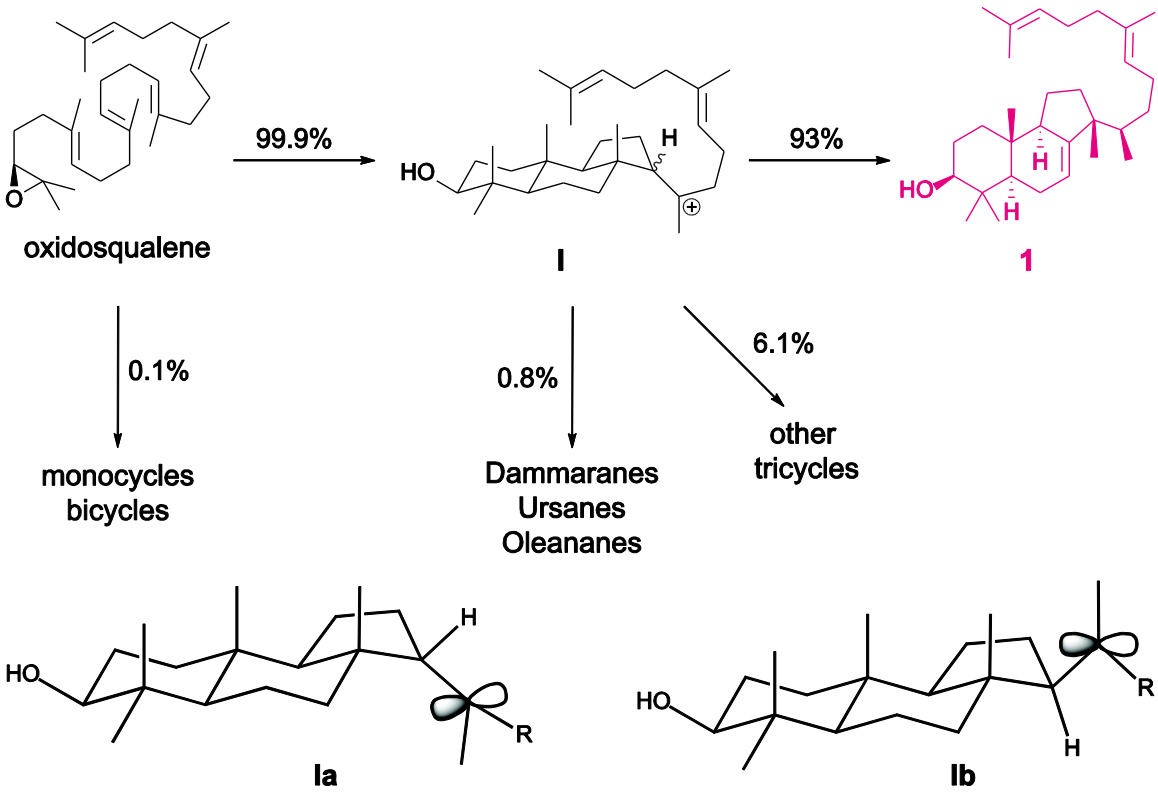


Figure 3.6. Comparison of active site residues in human LSS[7] and corresponding residues of *A. lyrata* OSCs.

3.3.2 Product Profile of *A. lyrata* PEN4

The product profile of *A. lyrata* PEN4[8] was determined by HSQC and GC-MS mixture analysis as described in Shan, H. *et al.*[9] and Castillo *et al.*[22] Ratios for the diols **2**, **10** and **14** were derived from the HSQC spectrum of the crude extract of the NSL, while the ratios for the triterpene monoalcohols were obtained from the HSQC of EHY41[pMMG13.2] SPE Fractions 22-35. Analysis of HPLC fractions by GC-MS was used to confirm the presence of any given product as part of the product profile, particularly minor products. Scheme 3.1 shows an overview of the formation of thalianol **1** and how oxidosqualene is distributed between the product profiles of *Aly*PEN4.



Scheme 3.1. Mechanism of formation of thalianol **1**, the major product of *Aly*PEN4. The role of intermediate cations **Ia** and **Ib** in product formation is discussed in Section 3.4.2 (Mechanism and Product Profile of *A. lyrata* PEN4).

Table 3.5 presents a comprehensive picture of the product profile of *Aly*PEN4 and its ortholog in *Ath*PEN4. Figure 3.7 shows a bar graph comparison between the minor products of *A. lyrata* and *A. thaliana* PEN4. Compound numbering follows the *A. lyrata* PEN4 product profile, from highest to lowest amounts.

Table 3.5. Summary of *Aly*PEN4 thalianol synthase products identified in the 4 L *in vivo* experiment.

Product #	Product	*A.lyrata* %amount	*A.thaliana* %amount	Group*	Ref**
1	thalianol	93	92	T	10
2	14-epiarabidiol	2	2	T	8
3	25-nor-9β-methylpodioda-5(10),17*E*,21-trien-3β-ol	1.6	1.8	T	(a)
4	14-epithalianol	0.9	1.1	T	8
5	Δ7-14-epithalianol	0.8	0.8	T	10
6	13αH-malabarica-14*Z*,17*E*,21-trien-3β-ol	0.5	0.6	T	10
7	13αH-malabarica-14*E*,17*E*,21-trien-3β-ol	0.2	0.5	T	10
8	camelliol C	0.1	0.2	M	11
9	tirucallol-8,24-dienol	0.08	0.09	D	13
10	(20*S*)-dammar-24-ene-3β,20-diol	0.05	0.08	D	14
11	9αH-polypoda-7,13*E*,17*E*,21-tetraen-3β-ol	0.05	0.03	B	(b)
12	13αH-malabarica-14(27)-trien-3β-ol	0.04	0.16	T	14
13	polypoda-8,13*E*,17*E*,21-tetraen-3β-ol	0.03	0.1	B	11
14	arabidiol	0.03	0.07	T	11
15	13βH-malabarica-14(27)-trien-3β-ol	0.02	0.04	T	14
16	dammara-20,24-dien-3β-ol	0.02	0.02	D	15
17	13βH-malabarica-14*E*,17*E*,21-trien-3β-ol	0.01	0.04	T	(c)
18	dammara-20(22)*E*,24-dien-3β-ol	0.01	0.02	D	16
19	9αH-polypoda-8(26),13*E*,17*E*,21-tetraen-3β-ol	0.01	0.01	B	17
20	tirucalla-7,24-dien-3β-ol	0.01	0.01	D	13
21	germanicol	0.01	0.01	O	18
22	taraxasterol	0.01	0.01	U	19

23	β-amyrin	0.01	0.01	**O**	20
24	α-amyrin	0.01	0.01	**U**	21
25	ψ-taraxasterol	0.01	0.01	**U**	19
26	Δ7-thalianol	<0.01	0.15	**T**	10
27	13βH-malabarica-14*Z*,17*E*,21-trien-3β-ol	<0.01	0.03	**T**	17

* **M: monocycles B: bicycles T: tricycles D: dammaranes U: ursanes O: oleananes**

** References: (a) No reference exists because neither C14 epimer of compound **3** is known. (b) Nguyen, L. H. D.; Harrison, L.J. *Phytochemistry* **1998**, *50*, 471-476. (c) Kolesnikova, M. D., Ph.D. Thesis, Rice University, 2008.

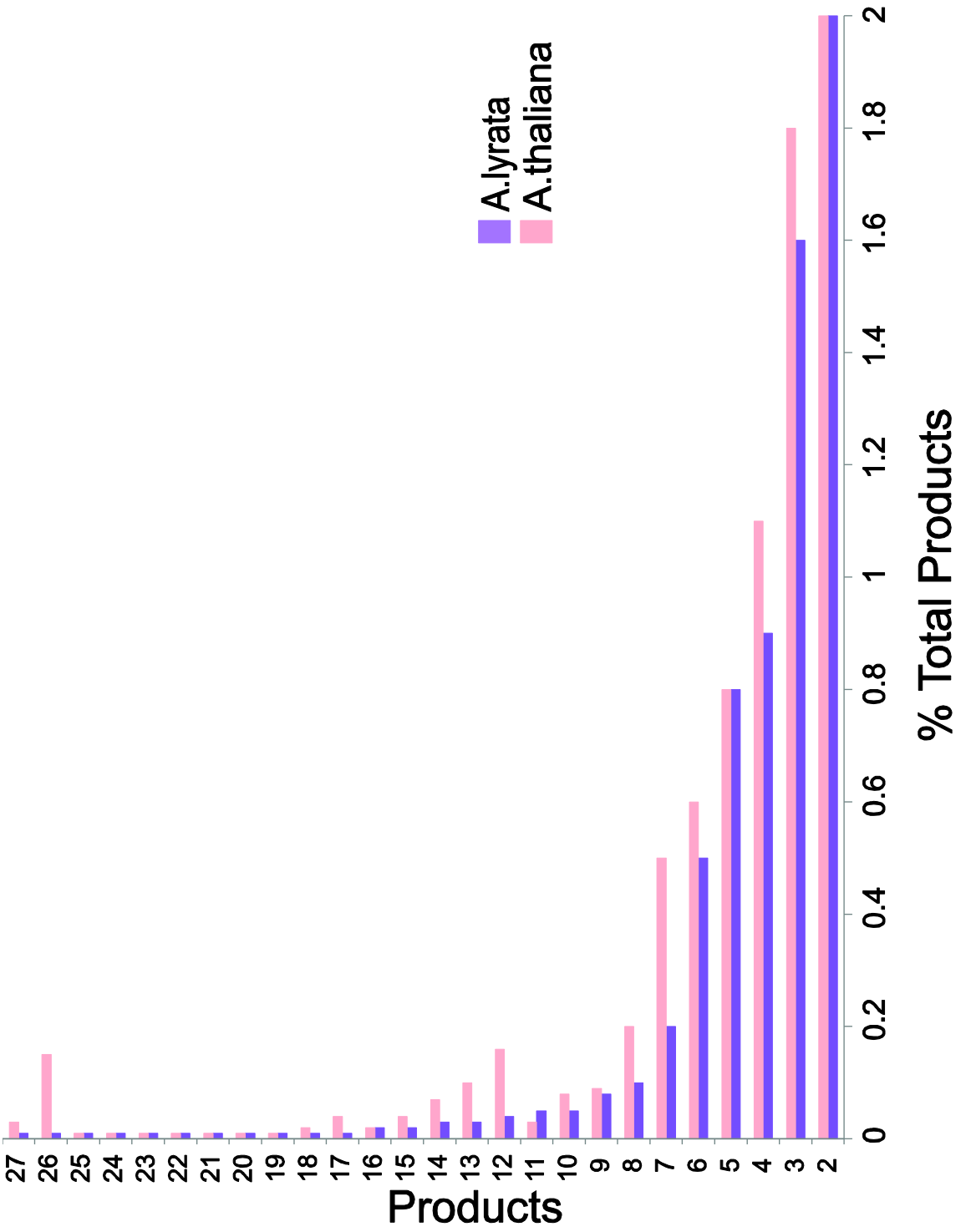


Figure 3.7. Product profile comparison between minor products of *A. lyrata* and *A. thaliana*. Violet bars represent the product profile of *Aly*PEN4 and pink bars represent the product profile of *Ath*PEN4.

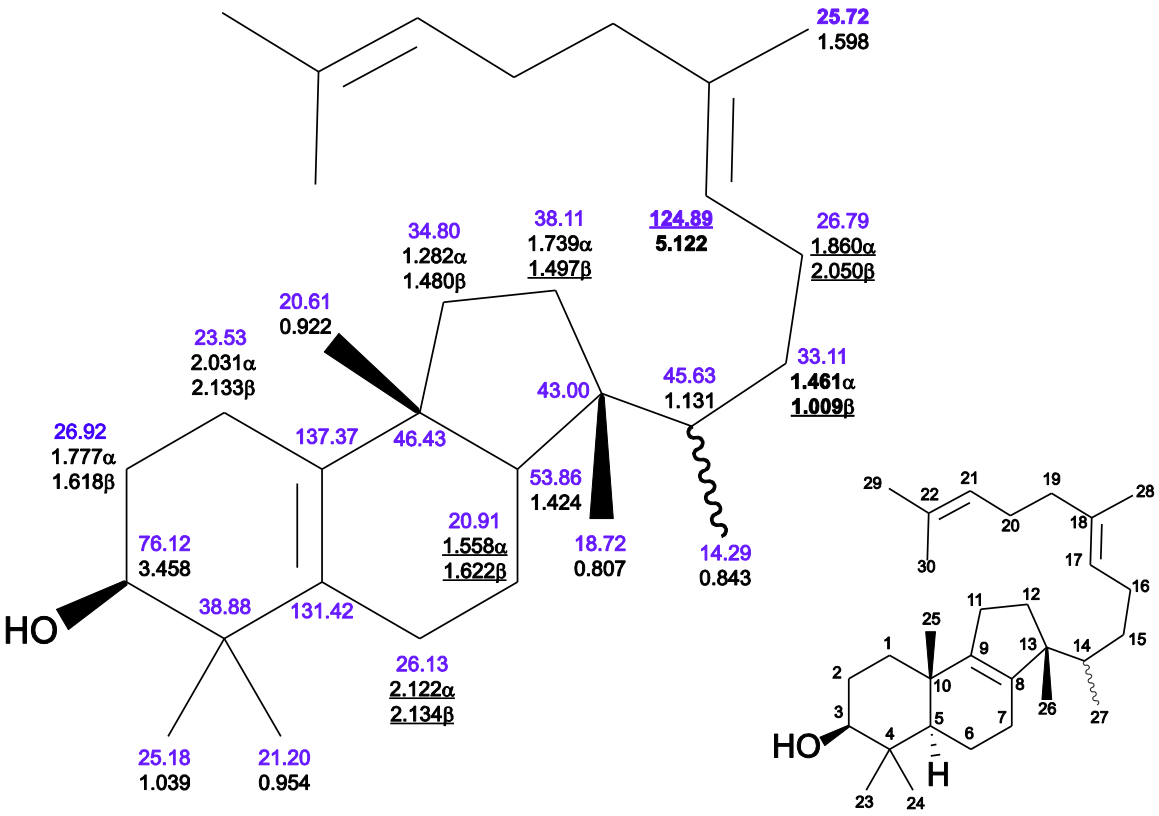


Figure 3.8. NMR assignments and atom numbering for 25-nor-9β-methylpodioda-5(10),17*E*,21-trien-3β-ol **3**.

*Aly*PEN4 third major product 25-nor-9β-methylpodioda-5(10),17*E*,21-trien-3β-ol **3** is a triterpene alcohol previously unknown. Its structure was determined on the course of this work.

This compound was partially isolated from the 4 L EHY41 triterpene mixture by reversed-phase HPCL analysis. The partial purification from preparative reversed-phase HPLC gave a fraction (Fx 43) that contained a 3:3:1 ratio of Δ7-thalianol epimer **5**, 25-nor-9β-methylpodioda-5(10),17*E*,21-trien-3β-ol **3**, thalianol **1**, and other products. ^{1}H NMR, HSQC, HMBC, and COSYDEC spectra were acquired for this mixture and used to determine C-C and C-H connectivities in order to assign the chemical shifts for this structure (Figure 3.8). The chemical shifts shown in Figure 3.8 are estimated as accurate to ±0.001 ppm for ^{1}H and ±0.01 ppm for ^{13}C except for underlined values (±0.003 ppm and

±0.03 ppm) and ^{1}H values given to two decimal places (±0.01 ppm). Chemical shifts are not corrected for strong coupling.

Quantum mechanical calculations (performed by William K. Wilson) were done in order to confirm experimental chemical shifts and study the C14 configuration. The calculations were done as described by Shan, H. *et al.*[9] and Castillo *et al.*[22] on a 25-carbon model structure shown in Figure 3.9. GIAO chemical shifts were calculated at the B3PW91/6-311G(2d,p)//B3LYP/6-31G* level for 4 side chain conformers of the 14*R* epimer and 19 conformers of the 14*S* epimer.

Preliminary data for observed and calculated chemical shifts are compared in Table 4.3 and Table 4.4. The calculated chemical shift for 25-nor-9β-methylpodioda-5(10),17*E*,21-trienol **3** validates the initial assignments of the observed chemical shifts for ^{13}C and ^{1}H.

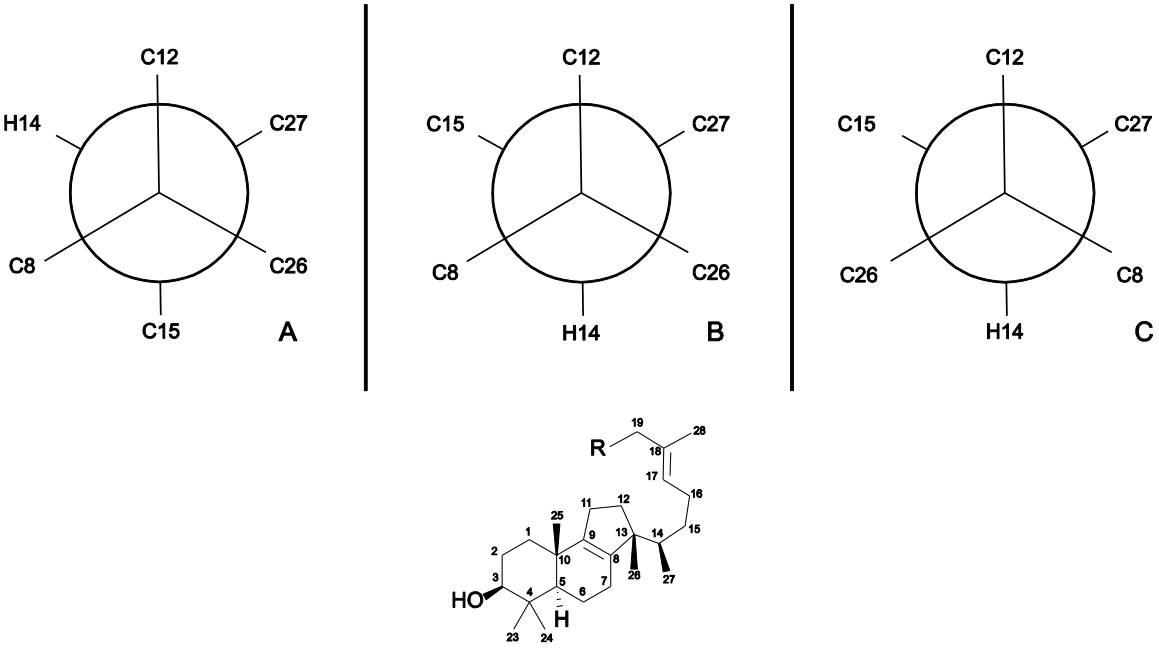


Figure 3.9. C13-C14 rotamers for the C_{25} model of 25-nor-9β-methylpodioda-5(10),17*E*,21-trien-3β-ol used for quantum mechanical modeling calculations. The 14*S* configuration is shown with R=H.

Table 3.6. Comparison of the observed and calculated ^{13}C NMR chemical shifts for **3**.

^{13}C atom	Obs δ	Calc δ	Δ Obs δ vs Calc δ
C-1	23.53	24.8	1.3
C-2	26.92	28.1	1.2
C-3	76.12	75.6	-.6
C-4	38.88	38.9	.0
C-5	131.42	132.8	1.4
C-6	26.13	27.0	.9
C-7	20.91	21.6	.7
C-8	53.86	54.5	.6
C-9	46.43	46.9	.4
C-10	137.37	137.1	-.3
C-11	34.80	33.6	-1.2
C-12	38.11	38.6	.5
C-13	43.00	43.6	.6
C-14	45.63	47.2	1.6
C-15	33.11	32.4	-.7
C-16	26.79	27.6	.9
4α-Me	25.18	24.9	-.3
4β-Me	21.20	20.8	-.4
9β-Me	20.61	20.8	.2
13β-Me	18.72	18.8	.0
14-Me	14.29	16.3	2.0
		average deviation	0.43
		rms deviation	0.91

These data are from 19 side-chain conformers of the 14*S* epimer. The conformers were side-chain variants of the rotamers shown in Figure 3.9. Because the set of side-chain conformers may be incomplete, these results are very preliminary. The separate corrections for the C4-C3-O-H rotamer heterogeneity are also preliminary, perhaps accounting for the substantial chemical shift deviations in ring A. The calculations for the 14*R* epimer (not shown) give greater deviations in the vicinity of C14 than for the 14*S* epimer. However, this may be due to the grossly incomplete set of conformers of the 14*R* epimer. Also, the fraction 43 sample should be purified and studied by NOESY as part of the determination of the C14 configuration. These calculations do strongly support the structure and NMR assignments for the carbons in the tricyclic ring system and nearby side-chain carbons.

Table 3.7. Comparison of the observed and calculated ^{1}H NMR chemical shifts for **3**.

^{1}H atom	Obs δ	Calc δ	Δ Obs δ vs Calc δ
H-1a	2.031	2.13	.09
H-1b	2.133	2.20	.01
H-2a	1.777	1.82	.05
H-2b	1.618	1.59	-.02
H-3a	3.458	3.53	.07
H-6a	2.122	2.24	.12
H-6b	2.134	2.23	.10
H-7a	1.558	1.61	.05
H-7b	1.622	1.62	.00
H-8a	1.424	1.67	.25
H-11a	1.282	1.39	.10
H-11b	1.480	1.56	.08
H-12a	1.739	1.79	.05
H-12b	1.497	1.60	.11
H-14	1.131	1.31	.18
H-15a	1.461	1.44	-.02
H-15b	1.009	1.03	.02
H-16a	1.860	1.76	-.10
H-16b	2.050	2.23	.18
4α-Me	1.039	1.11	.07
4β-Me	.954	.97	.02
9β-Me	.922	.98	.06
13β-Me	.807	.87	.06
14-Me	.843	.75	-.09
		average deviation	0.06
		rms deviation	0.095

See notes for Table 3.6.

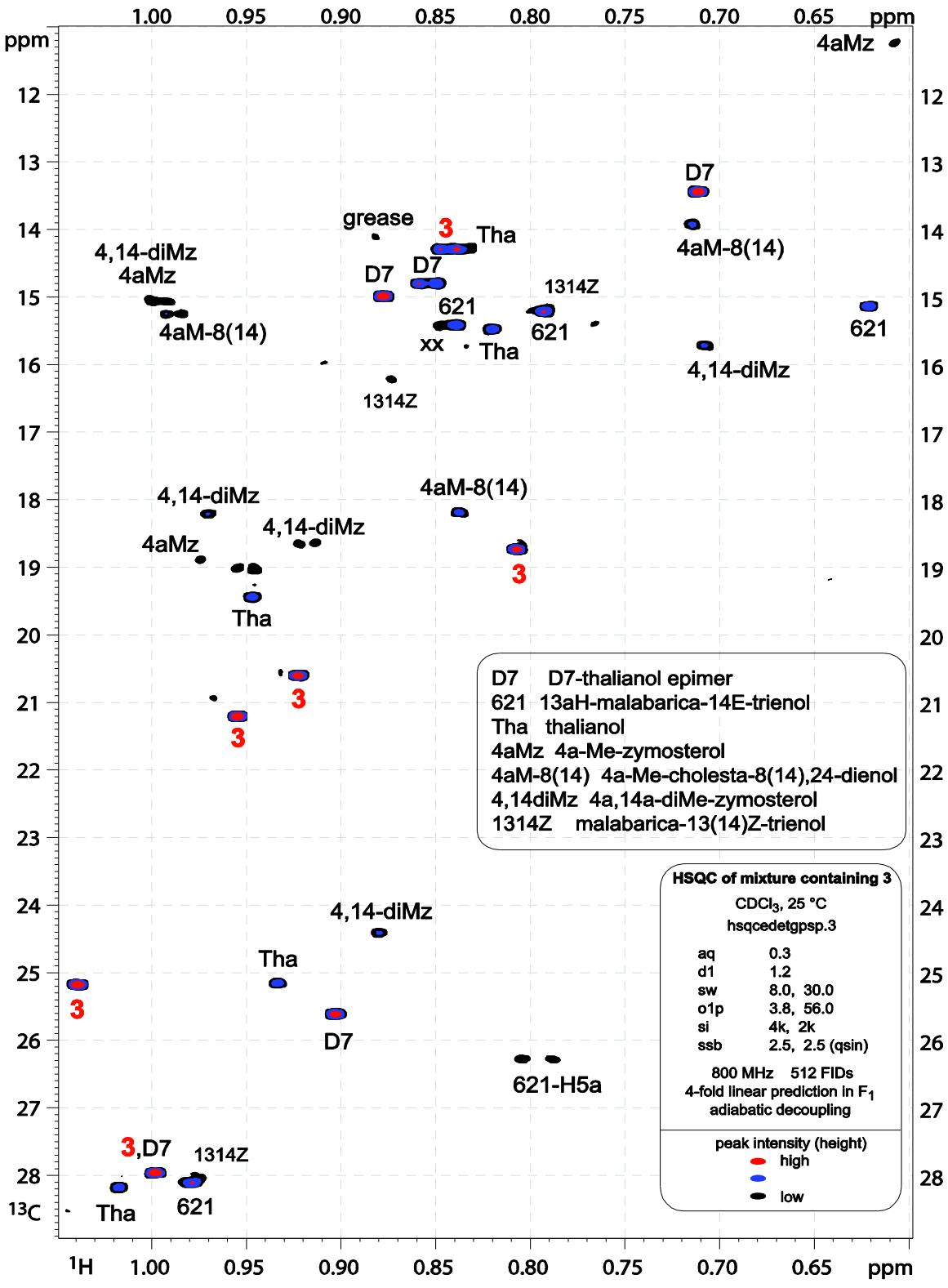


Figure 3.10. HSQC spectrum of HPLC fraction 43, a mixture containing 14-epi-Δ7-thalianol **5**, 25-nor-9β-methylpodioda-5(10),17*E*,21-trien-3β-ol **3**, and thalianol **1** (3:3:1 ratio) and other products.

3.3.3 Product Profile of *A. lyrata* PEN6

A. lyrata PEN6 is an ortholog to *A. thaliana* PEN6.[24] The major product seco-β-amyrin **1**, accounts for the 72% of the total of products. In addition to seco-β-amyrin **1**, this cyclase makes at least 39 other triterpene alcohols.

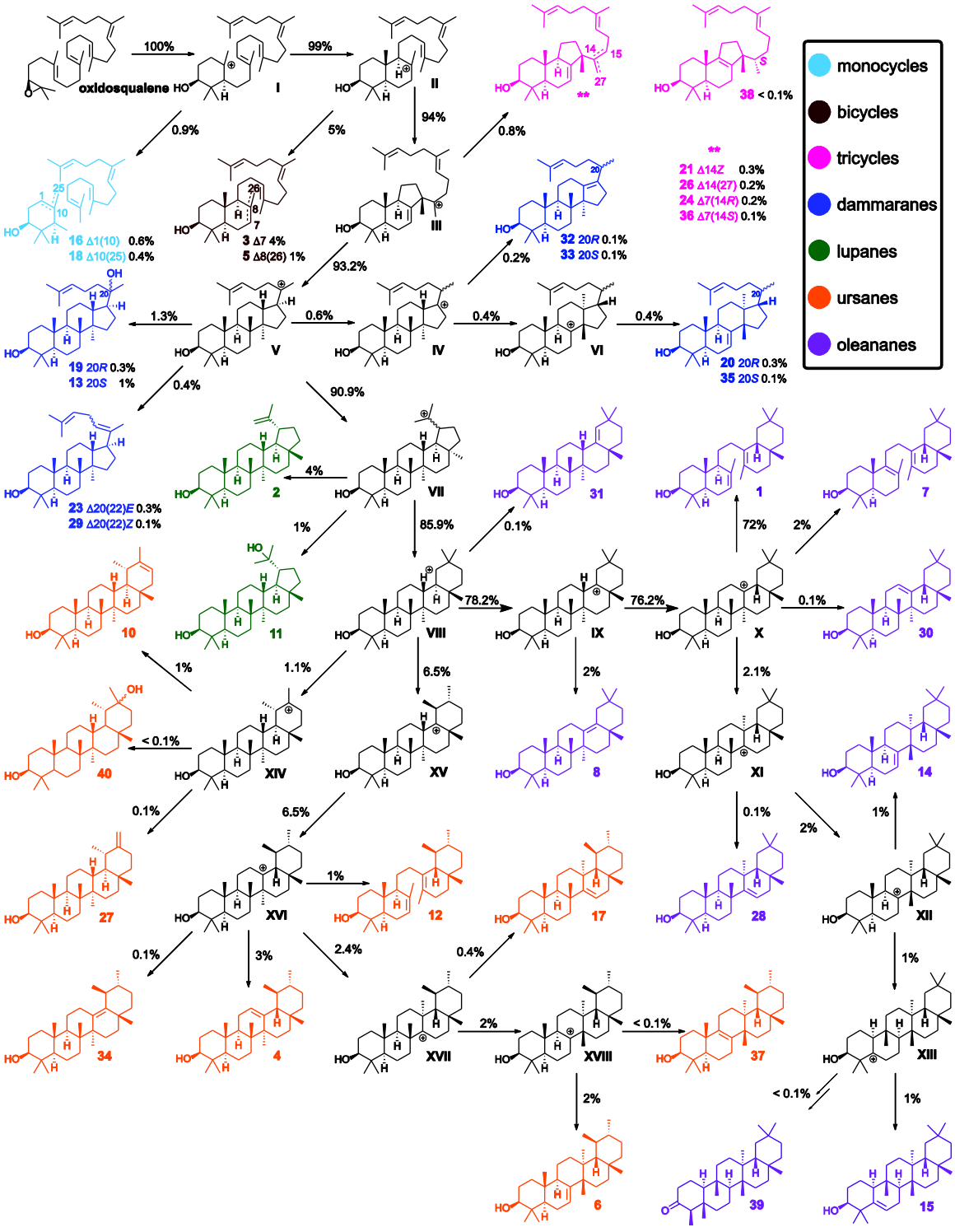
The product profile was determined using HSQC mixture analysis as described in Shan, H. *et al.*[9] Ratios for the diols **13**, **19** and **40** were determined from the HSQC of the crude extract of NSL, while the ratios for the triterpene alcohols were determined using the HSQC spectrum of EHY41[pMMG14.2] SPE Fx21-35. GC-MS analysis of HPLC fractions to confirmed the presence of each product, particularly minor products. Table 3.8 presents a comprehensive list of the products of *Aly*PEN6 and its ortholog *Ath*PEN6, and a proposed mechanistic pathway to each product is presented in Scheme 3.2.

Figure 3.11 shows a bar graph comparison between the minor products of *A. lyrata* PEN6 and *A. thaliana* PEN6.[18,23] Compound numbering follows the *A. lyrata* PEN6 product profile, from highest to lowest amounts and is independent of the PEN4 and PEN9 numbering.

Table 3.8. Summary of *Aly*PEN6 seco-β-amyrin synthase products identified in the 4 L *in vivo* experiment.

Product #	Product	*A.lyrata* % amount	*A.thaliana* % amount	Ref
1	seco-β-amyrin	72	38	23
2	lupeol	4	3	24
3	polypoda-7,13*E*,17*E*,21-tetraen-3β-ol	4	5	17
4	α-amyrin	3	3	21
5	9αH-polypoda-8(26),13*E*,17*E*,21-tetraen-3β-ol	2	0.3	17
6	bauerenol	2	6	25
7	Δ8(26)-seco-β-amyrin	2	0.1	17
8	δ-amyrin	2	1	26
9	dammara-20,24-dien-3β-ol	1	6	15
10	ψ-taraxasterol	1	1	19
11	lupane-3β,20-diol	1	2	27

12	seco-α-amyrin	1	19	23
13	(20*S*)-dammar-24-ene-3β,20-diol	1	2	14
14	multiflorenol	1	1	28
15	glutinol	1	< 0.1	29
16	camelliol C	0.5	0.5	11
17	isoursenol	0.4	5	30
18	achilleol	0.4	0.2	11
19	(20*R*)-dammar-24-ene-3β,20-diol	0.3	2	14
20	butyrospermol	0.3	0.3	15
21	13βH-malabarica-14*Z*,17*E*,21-trien-3β-ol	0.3	2	17
22	nematocyphol	0.3	0.3	31
23	dammara-20(22)*E*,24-dien-3β-ol	0.3	0.2	16
24	Δ7-thalianol	0.2	0	--
25	13αH-malabarica-14*E*,17*E*,21-trien-3β-ol	0.2	0	10
26	13βH-malabarica-14(27)-trien-3β-ol	0.2	1	14
27	taraxasterol	0.1	1	19
28	taraxerol	0.1	0.4	32
29	dammara-20(22)*Z*,24-dien-3β-ol	0.1	0.1	16
30	β-amyrin	0.1	0	20
31	germanicol	0.1	0	18
32	isoeuphol	0.1	0.1	33
33	isotirucallol	0.1	0.1	33
34	γ-amyrin	0.1	0.1	34
35	tirucalla-7,24-dien-3β-ol	0.1	< 0.1	12
36	Δ7-14-epithalianol	0.1	0	--
37	isobauerenol	< 0.1	< 0.1	35
38	14-epithalianol	< 0.1	0	8
39	friedelin	0	0.1	29
40	(20*R*)-taraxastane-3β,20-diol	0	< 0.1	36



Scheme 3.2. Proposed mechanistic pathways leading to PEN6 product profile.

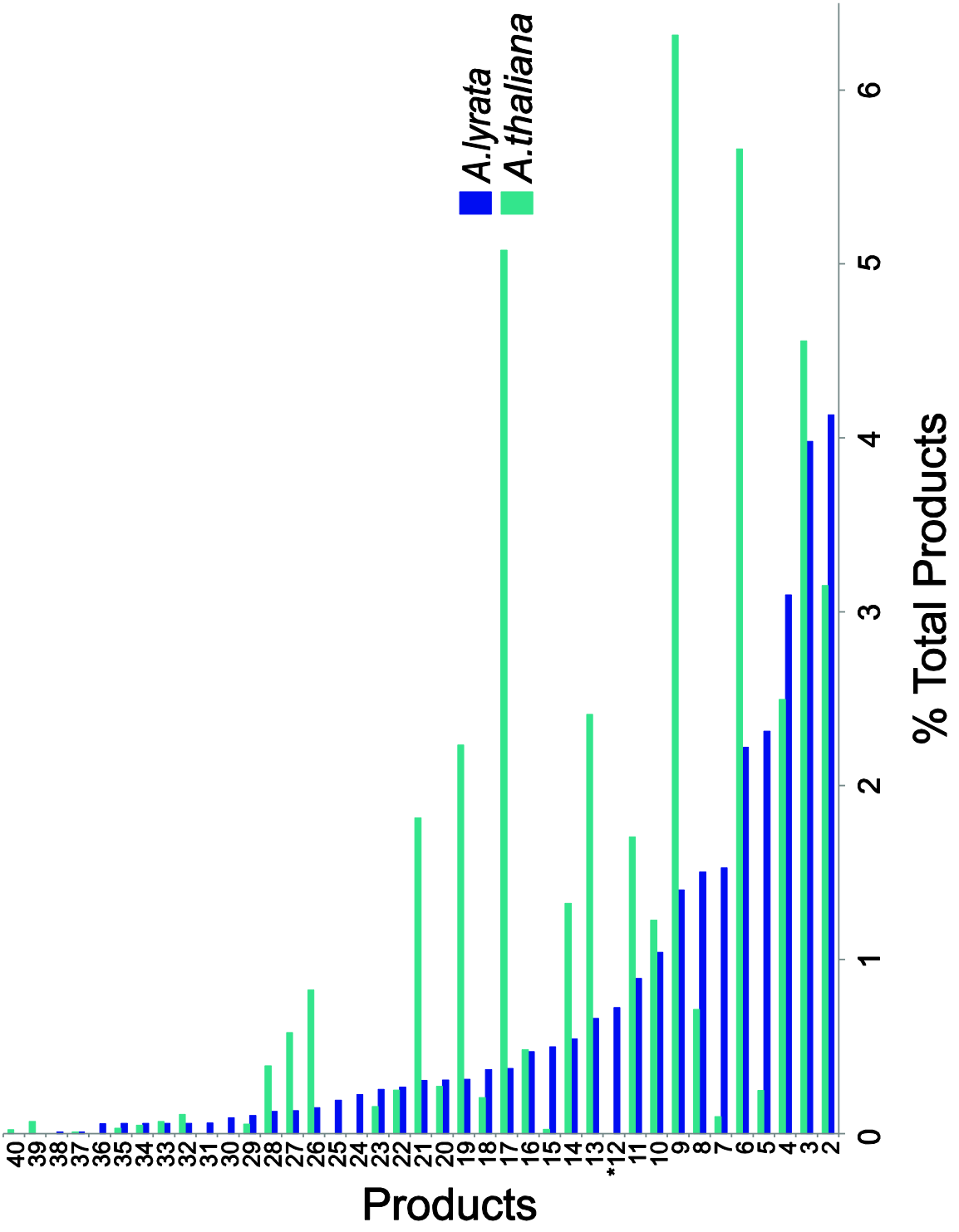


Figure 3.11. Product profile comparison between minor PEN6 products of *A. lyrata* and *A. thaliana*. Blue bars represent the product profile of *Aly*PEN6 and green bars represent the product profile of *Ath*PEN6.[17, 23] Product **12***, which is the second major product in *A. thaliana*, is not shown so that minor products could be visible.

3.3.4 *A. lyrata* seeds plant extracts

Table 3.9 shows a detailed summary of the triterpene alcohols found in *A.lyrata* seeds from Folch extraction and saponification. These results were obtained from HSQC of the NSL.

Table 3.9. Summary of names, numbering and percentage amounts for compounds found in *A. lyrata* seeds extracts.

Product # *	Product name	*A. lyrata* seed extracts, % total	*Aly*PEN6 % total
1	seco-β-amyrin	71	71
2	lupeol	4	2
3	9αH-polypoda-7,13*E*,17*E*,21-tetraen-3β-ol	4	3
4	α-amyrin	3.1	1.8
5	9αH-polypoda-8(26),13*E*,17*E*,21-tetraen-3β-ol	2.3	1.7
6	bauerenol	2.2	1.3
7	Δ8(26)-seco-β-amyrin	1.5	1.6
8	δ-amyrin	1.5	0.8
9	dammara-20,24-dien-3β-ol	1.4	1.1
10	ψ-taraxasterol	1.2	0.7
11	lupane-3β,20-diol	0.9	0.7
12	seco-α-amyrin	0.7	0.7
13	(20*S*)-dammar-24-ene-3β,20-diol	0.7	0.7
14	multiflorenol	0.5	0.6
15	glutinol	0.5	0.7

* Compound numbering follows the product profile of *Aly*PEN6.

The table above also shows the comparison between the ratios of the triterpene product profiles of *Aly*PEN6 and the seeds extracts from *A. lyrata*. The ratios are remarkably similar, suggesting that the PEN6 products undergo little or no further metabolism. Only minor amounts of other secondary triterpene metabolites were detected. Thus, PEN6 is essentially the only OSC besides CAS that contributes to seed triterpenes.

The results also support the assumption that yeast heterologous expression of OSCs follows the product profile in plants.

3.3.5 Product Profile of *A. lyrata* PEN9

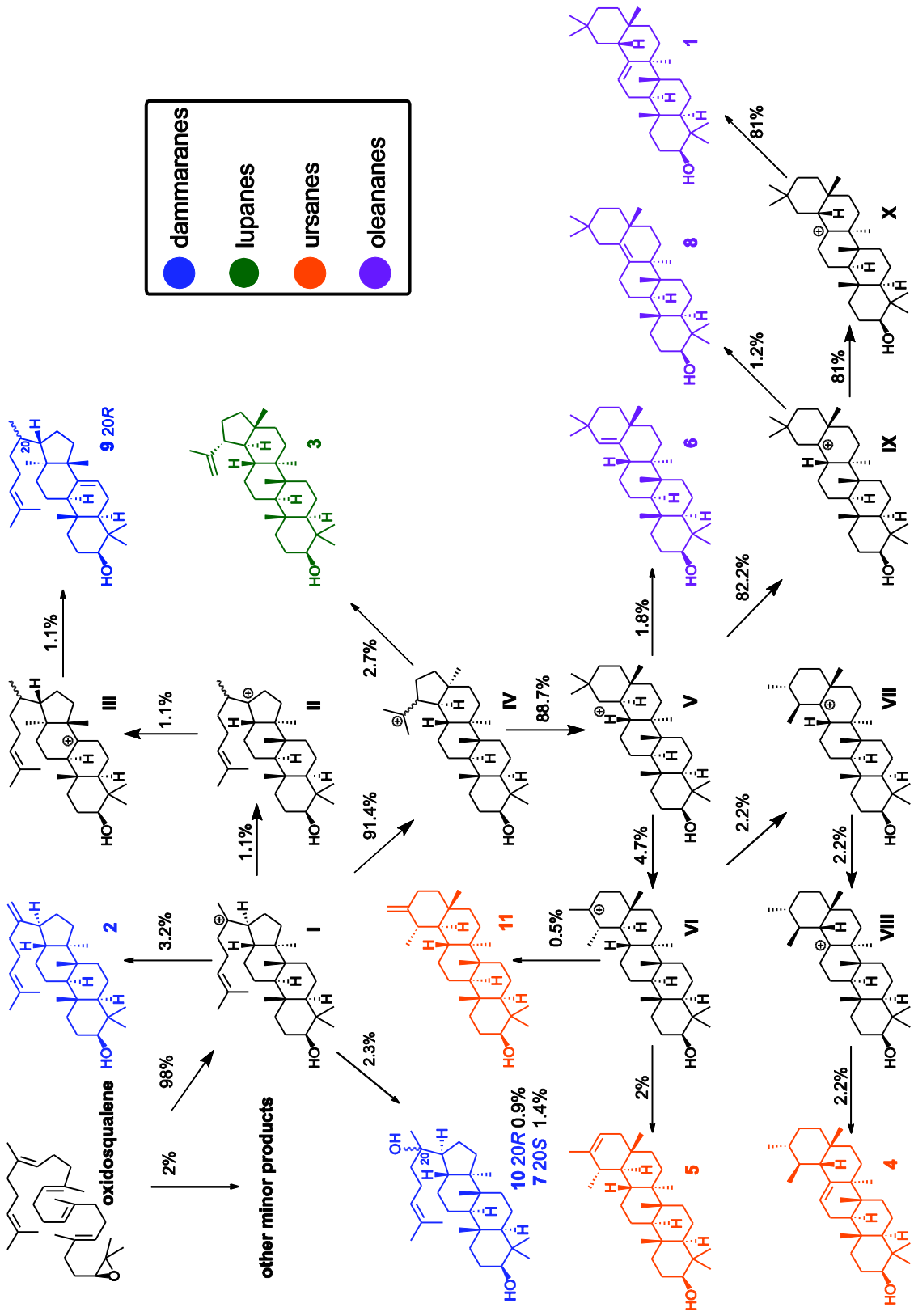
Table 3.10. Summary of *Aly*PEN9 β-amyrin synthase products identified in the 4 L *in vivo* experiment.

Product #	Product	*A. lyrata* % amount	Reference
1	β-amyrin	81	20
2	dammara-20,24-dien-3β-ol	3.2	15
3	lupeol	2.7	24
4	α-amyrin	2.2	21
5	ψ-taraxasterol	2.0	19
6	germanicol	1.8	18
7	(20*S*)-dammar-24-ene-3β,20-diol	1.4	13
8	δ-amyrin	1.2	26
9	butyrospermol	1.1	15
10	(20*R*)-dammar-24-ene-3β,20-diol	0.9	19
11	taraxasterol	0.5	27
*	minor products	2	na

* Minor products: taraxerol, tirucalla-7,24-dien-3β-ol, dammara-20(22)*Z*,24-dien-3β-ol, γ-amyrin, isoephol, dammara-20(22)*E*,24-dien-3β-ol, isotirucallol, multiflorenol.

A. lyrata PEN9 does not have an ortholog in the *A. thaliana* genome. The major product β-amyrin **1**, accounts for 81% of the total of products. In addition to **1**, this cyclase makes at least 18 other triterpenes.

The enzyme mechanism and product profile are presented in Scheme 3.3. The product profile was determined as described in Shan, H. *et al.*[9] Ratios for the diols **7** and **10** were determined from the HSQC of the crude extract of NSL, while the ratios for the triterpene alcohols were determined using HSQC of EHY41[pMMG9.1] SPE Fx 23-30. Analysis of HPLC fractions by GC-MS was used to confirm the presence of any given triterpene as part of the product profile, particularly minor products. Table 3.10 presents a detailed picture of the product profile of this enzyme.



Scheme 3.3. Proposed mechanistic pathways leading to PEN9 product profile.

3.4 Discussion

3.4.1 Phylogenetic Analysis of OSCs in *A. lyrata* and *A. thaliana*

The 14 putative OSCs of *A. lyrata* are shown in three phylogenetic groups (Figure 3.12). Further analysis of genomic DNA, protein sequences, active site amino acids residues (Figure 3.13), and plant extracts indicated that *A. lyrata* has nine functional OSCs.

The phylogenetic analysis of *A. lyrata* OSCs was the first step to identify putative OSCs, for that reason the following phylogenetic discussion will treat pseudogenes as putative OSCs since sequence ambiguities were confirmed later on in the study.

A. lyrata OSCs can be grouped in three main categories: (1) those OSCs that share high overall sequence identity and have identical active site residues to their ortholog in *A. thaliana*, (2) those cyclases that have high sequence identity but some differences in active site residues, and (3) those which do not have an ortholog in the *A. lyrata* genome.

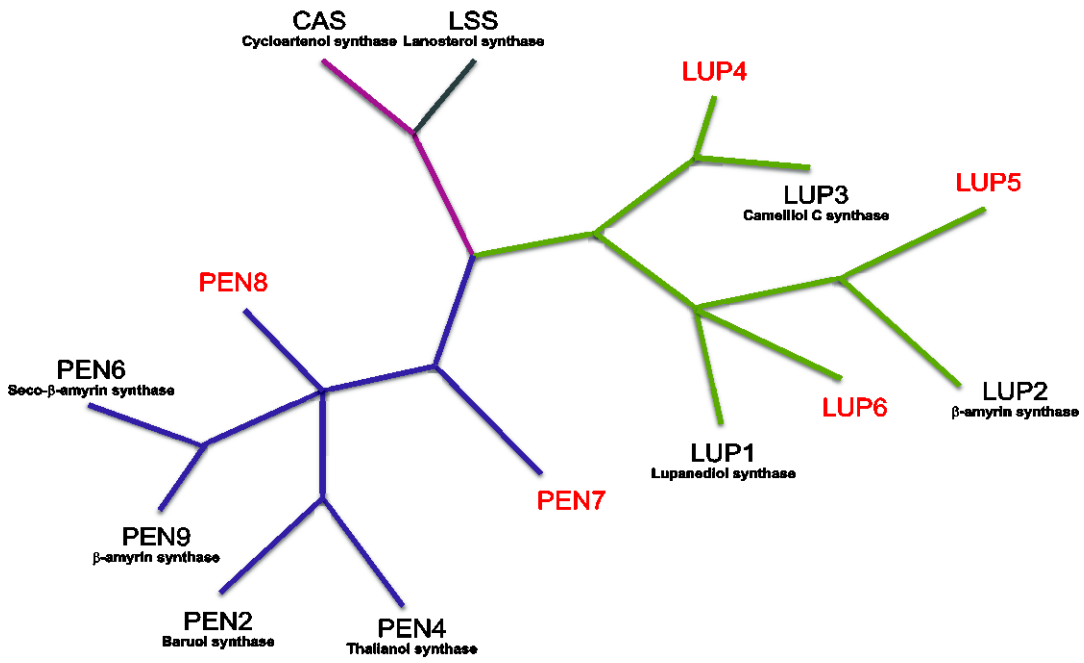


Figure 3.12. The 14 OSCs from *A. lyrata* are separated in three phylogenetic groups CAS (pink), LUP (green) and PEN (purple). Putative pseudogenes are shown in red.

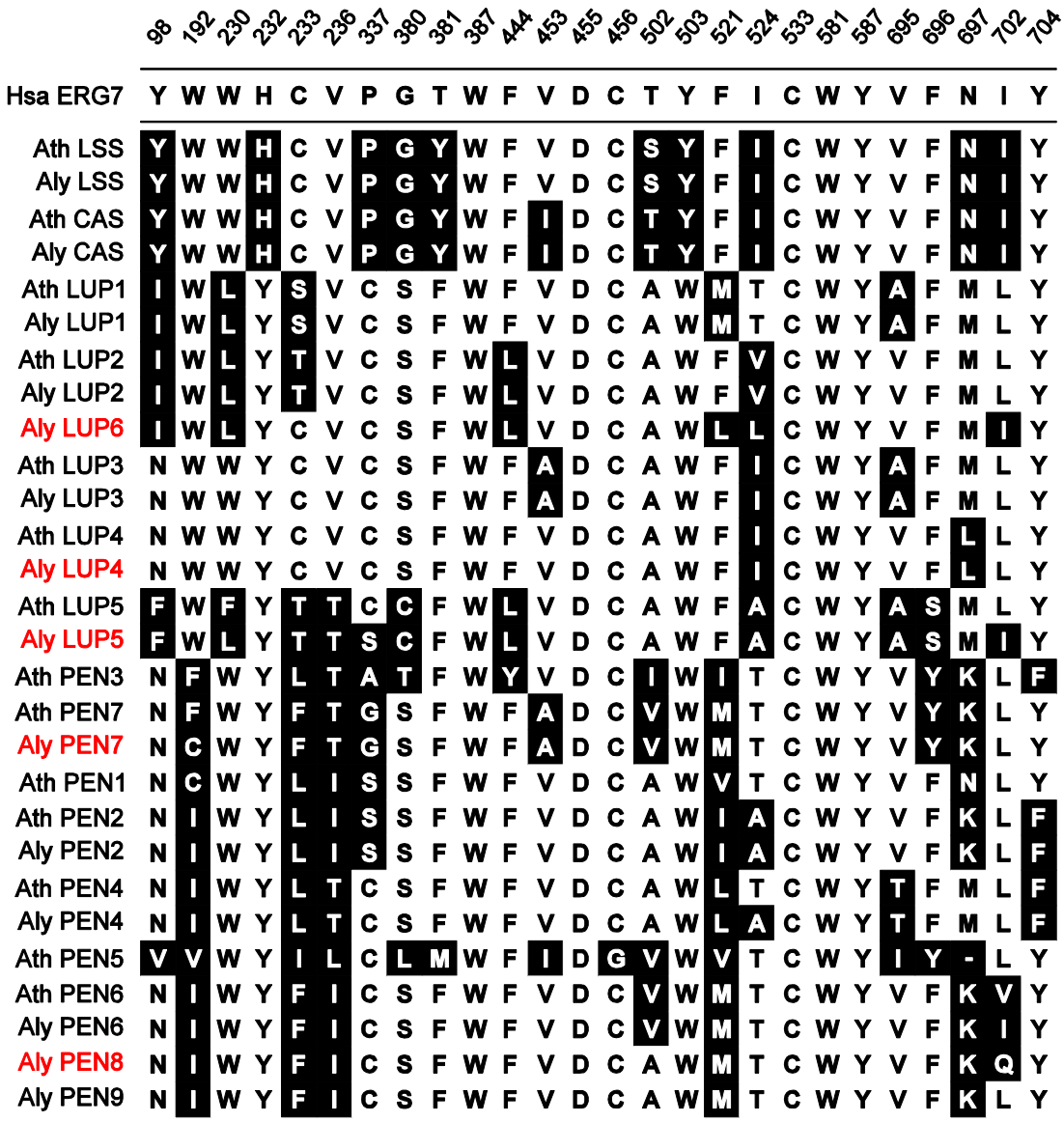


Figure 3.13. Comparison of active site residues in human LSS[7] and corresponding residues of *A. lyrata* and *A. thaliana* OSCs. Putative pseudogenes are shown in red.

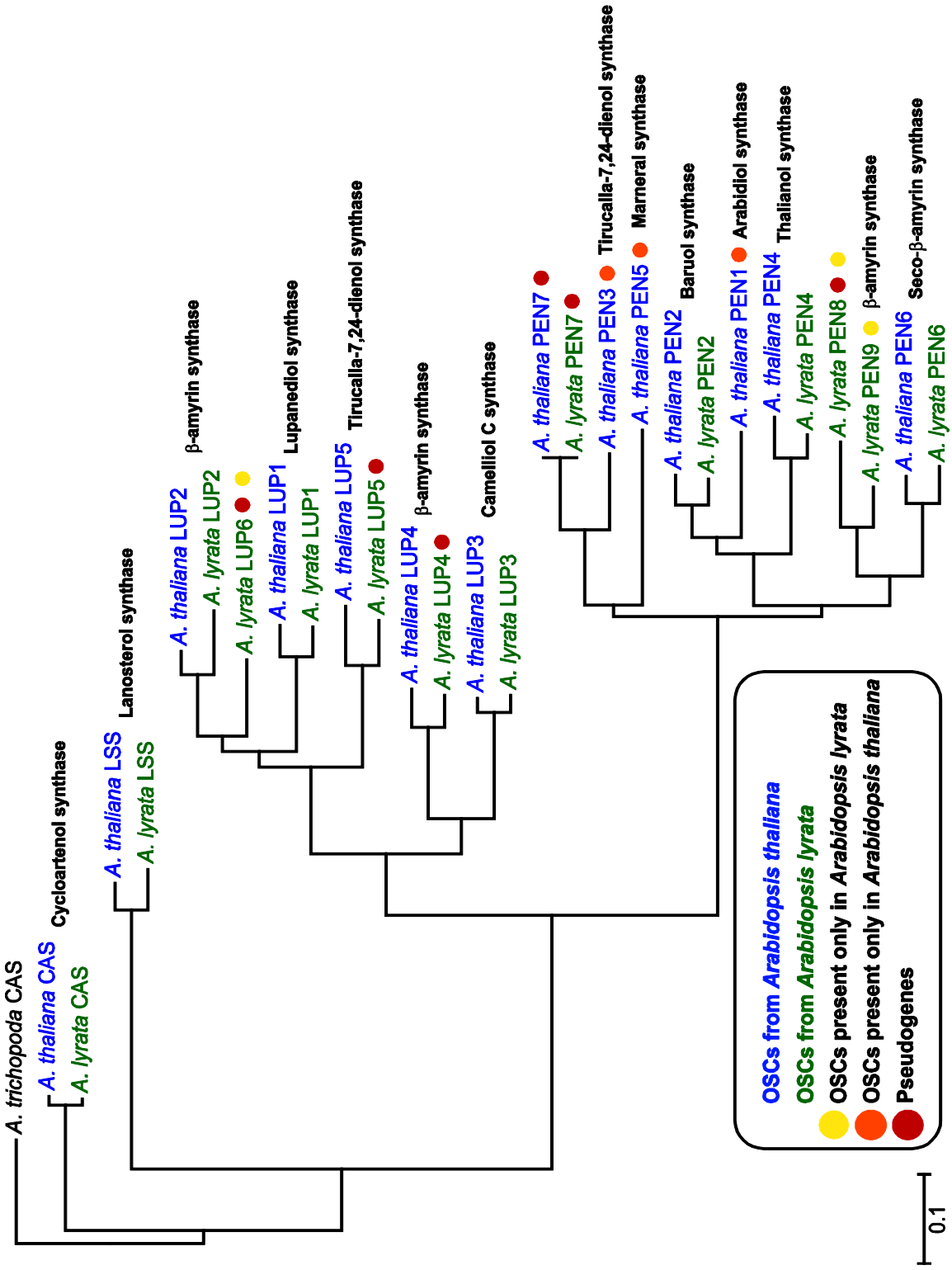


Figure 3.14. Phylogenetic tree of *A. lyrata* OSCs after manual splicing of the genome. The tree was constructed by maximum likelihood using MEGA version 6.[5] Sequences were aligned with Clustal W using default parameters.

The OSCs that share high sequence identity (in parenthesis) and identical active site residues with their *A. thaliana* orthologs are *Aly*CAS (98%), *Aly*LSS (93%), *Aly*LUP1 (93%), *Aly*LUP2 (92%), *Aly*LUP3 (95%), *Aly*PEN2 (94%). This high similarity strongly suggests that the product profile of these six enzymes would closely resembles the product profile of the already characterized OSCs from *A. thaliana*.

The phylogenetic tree comparison (Figure 3.14) shows that three *A. thaliana* OSCs are absent from the *A. lyrata* genome: arabidiol synthase (PEN1), tirucalla-7,24-dien-3β-ol synthase (PEN3) and marneral synthase (PEN5). Thus, *A. lyrata* does not produce these compounds or their P450 metabolites.[22,37]

*Aly*LUP5 (89%), *Aly*PEN4 (93%) and *Aly*PEN6 (93%) have high overall sequence identity to their orthologs in *A. thaliana*, but one or two differences in the active site residues. This will make them good candidates for expression (apart from pseudogene *Aly*LUP5) since changing an active site residue can change the catalytic outcome of the enzyme.[38]

Also uncovered by the phylogeny is that four genes that are absent in *A. thaliana*'s genome were identified in *A. lyrata*: *Aly*LUP6 (84% to *Ath* LUP2), *Aly*PEN8 (82% to *Ath*PEN6), *Aly*PEN9 (78% to *Ath*PEN6) and *Aly*PEN7 (ortholog to *Ath*PEN7 which is a pseudogene in *A. thaliana*). These novel OSCs are interesting candidates for study since their product profiles cannot be predicted reliably. However, note that *Aly*PEN7 and *Aly*PEN8 appear to be pseudogenes.

In the LUP clade (Figure 3.12) several splicing issues arose. The first disagreement was found in the splicing of the first exon of *Aly*LUP2 (92%). This OSC was manually spliced using as a template its ortholog in *A. thaliana*; the only disagreement was that the genomic DNA sequence from Phytozome[2] had 4 extra nucleotides at the end of the first exon. Genomic DNA sequence confirmed these nucleotides in the sequence, thus making a different splicing for *Aly*LUP2 first exon.

*Aly*LUP3 (95%) had a similar problem; an extra nucleotide at the end of the last exon created a longer exon for this protein, adding 14 amino acids at the end. Even though *Aly*LUP3 has this extra 14 amino acids it seems that this does not affect expression of the OSC in the plant.

A. lyrata plant extracts from flowers and bud contained camelliol and 29-hydroxycamelliol (previously found by Jing Jin in Capsella and later by Hui Shan and William Wilson in *A. thaliana* plant extracts). This information not only confirms the active presence of *Aly*LUP3 in the plant, but also suggests that the changes in the protein sequence do not affect the major product of the enzyme. *A. lyrata* and *A. thaliana* synthesize camelliol and its 29-hydroxy derivative when the bud and the flowers are being produced. This conservation of *Aly*LUP3 and the P450 during the 10 million year period since thaliana-lyrata divergence, when most ancestral OSCs were lost in at least one species, indicate an important role of these compounds in reproduction.

Finally genomic DNA analysis helped to determine whether *Aly*LUP5 (89% identical to *Ath*LUP5, with two changes in the active site, C337S and L702I), *Aly*LUP4 (93% identical to *Ath*LUP4, identical active site) and *Aly*LUP6 were pseudogenes. All of those cyclases had missing nucleotides within exons, problems that created frameshifts and thus shortened the length of the above OSCs significantly. Sequencing analysis supported my hypothesis that these OSCs are pseudogenes.

Compared with these LUP clade differences, the PEN clade of *A. lyrata* suffers more changes relative to *A. thaliana*. In this clade, we have the absence of three OSCs of the *A. thaliana* genome (PEN1, PEN3 and PEN5) and also additional putative OSCs specific to *A. lyrata* (PEN7, PEN8 and PEN9).

During the analysis of the genomic DNA sequence of *Aly*PEN7, it was found that the *Aly*PEN7 genomic sequence has an odd patch in which there are 2 nucleotides more than the *Ath*PEN7 genomic sequence. The Phytozome[2] database had that region spliced out so that the coding frame is maintained. However, doing so deletes the highly conserved EGGW sequence that corresponds to the DGGWGES sequence that is present in all expressed OSCs from *A. thaliana*.

*Aly*PEN7 was spliced in silico using *Ath*PEN3 as a reference. Both sequences were aligned and the introns were manually removed from the *Aly*PEN7 genomic sequence; the synthetic clone conserves the EGGW region. To prove whether the conflicted region has these extra two base pairs that create a frame, genomic DNA was extracted from *A. lyrata*. The conflict sequence was PCR amplified and sequenced. The results show that these extra nucleotides (A and T) are present in the genomic sequence of *A. lyrata*. The presence of

these two extra nucleotides in the exon with the DGGWGES sequence makes a frame shift that creates a stop codon. This evidence strongly suggests that *Aly*PEN7 is a pseudogene since the encoded protein is very short and lacks the EGGW region.

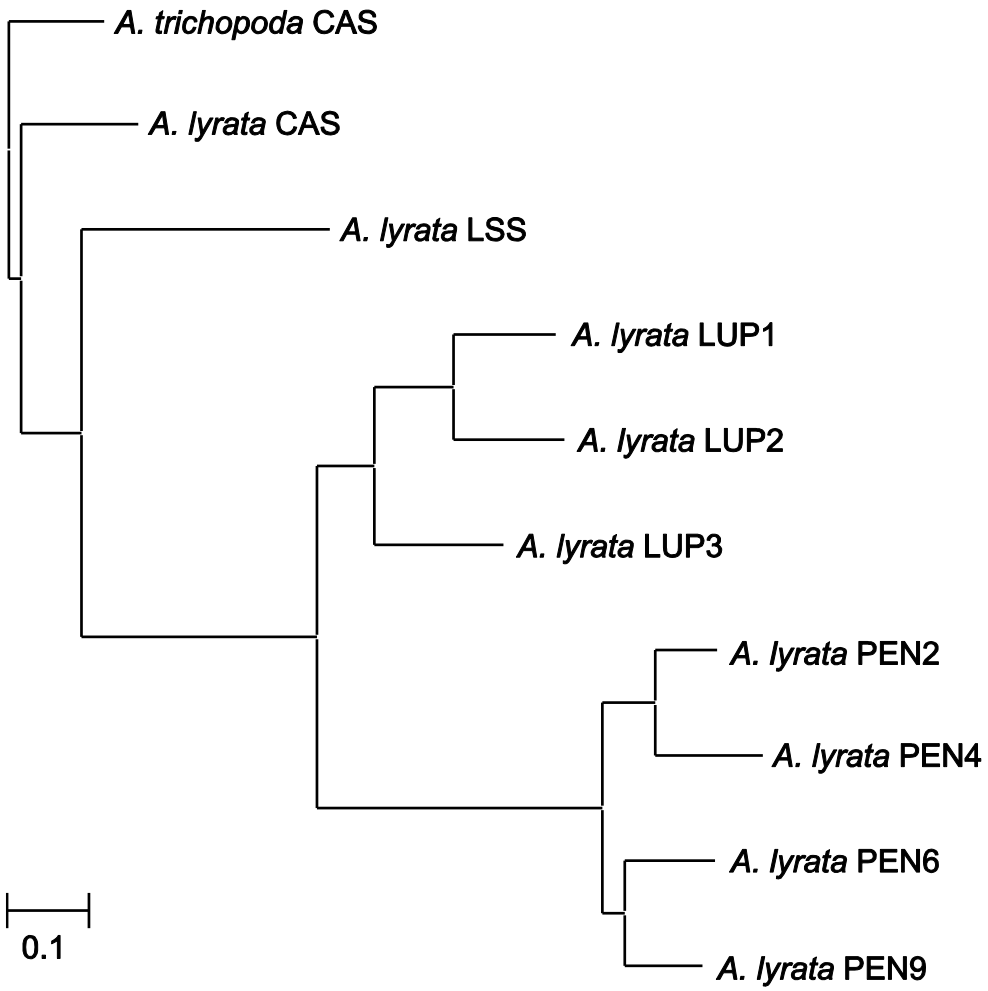


Figure 3.15. Molecular phylogenetic tree of *A. lyrata* OSCs found by manual splicing of the genome. The tree was constructed by maximum likelihood using MEGA version 6.[5] Sequences were aligned with Clustal W using default parameters.

The *A. lyrata* version of PEN8 that was available in Phytozome[2] was proofread by manual splicing using as a template its closest homolog *Ath*PEN6. The clone was compared against its genomic sequence and it was found that, what was previously spliced and expressed as *Aly*PEN8 was a chimera, i.e. *Aly*PEN8 from GeneArt contained the first exon of *Aly*PEN9. *Aly*PEN8 and *Aly*PEN9 are neighboring genes, so the computer splicing

missed an exon of *Aly*PEN9 which, upon a second round of comparison between genomic DNA and manually spliced genes, was found. Both PEN8 and PEN9 have no ortholog in *A. thaliana* and would be interesting to express. Figure 3.15 shows the final phylogenetic tree for *A. lyrata*.

Genomic DNA sequencing, manual splicing and active site amino acid comparison helped to determine which cyclases were likely to be orthologs to the *A. thaliana* OSCs. Thanks to the previous studies on the model organism *A. thaliana*, most of the triterpene product profile of *A. lyrata* can be guessed based on analysis of the genome, with minimal experimental work. However there still are questions to be answered regarding the product profile of the four putative cyclases. Heterologous gene expression was necessary to elucidate the product profile of *Aly*PEN8 and *Aly*PEN9, since neither has an ortholog in *A. thaliana*, and it was also important to express *Aly*PEN4 and *Aly*PEN6 cyclases because of the possibility of changes in the product profile since both of them exhibit differences in the amino acid sequence of the active site when compared with their *A. thaliana* orthologs (Figure 3.16 and Figure 3.17).

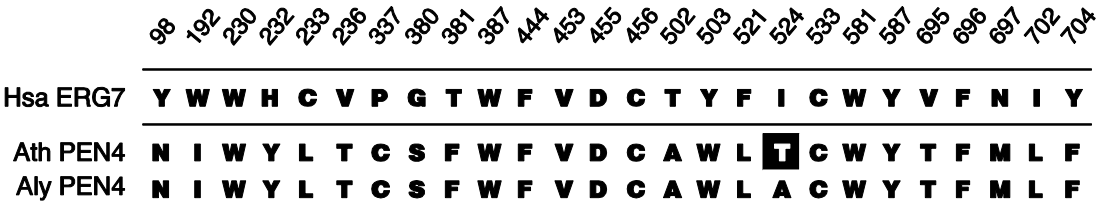


Figure 3.16. Comparison of active site residues of PEN4 orthologs.

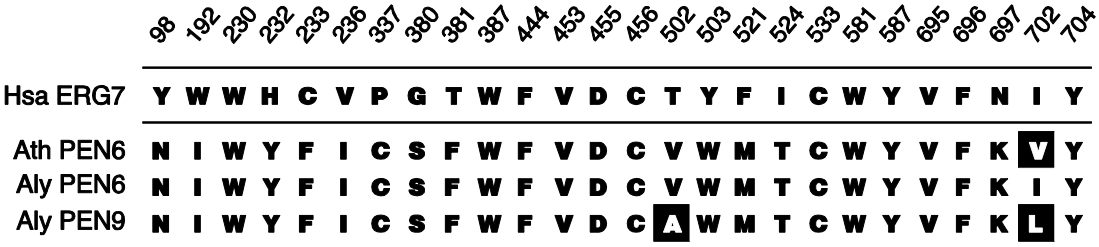


Figure 3.17. Comparison of active site residues of PEN6 orthologs.

*Aly*PEN4, *Aly*PEN6 and *Aly*PEN9 were successfully expressed in yeast; however expression of *Aly*PEN8 was not successful. *Aly*PEN8 is 82% identical to *Ath*PEN6 and looks like a full length OSC. Although *Aly*PEN8 was modified using PCR to introduce the correct "last exon" (not the chimeric version synthesize by GeneArt), several attempts at expression in yeast gave no triterpenes. Possibly *Aly*PEN8 has undergone genetic drift and is no longer a functional OSC. There is no EST library for *A. lyrata*, so there is no evidence of expression of this putative OSC in the plant. An attempt to obtain cDNA of this cyclase was made using different parts of the plant, but PCR amplification with *Aly*PEN8 specific primers did not show any results.

3.4.2 Mechanism and Product Profile of *A. lyrata* PEN4

A. lyrata PEN4 is an ortholog to *A. thaliana* PEN4. The sequence identity between orthologs is 93%. Despite one change among the active site residues, the product profile of the *A. lyrata* ortholog is almost identical. The mechanism of formation of the product profile and ratios comparison between *Aly*PEN4 and *Ath*PEN4 is illustrated in the results section Scheme 3.1, Table 3.5 and Figure 3.7.

The major product for both enzymes is thalianol **1**, previously characterized by Fazio *et al.*[8] This compound comes from protonation of OS follow by cyclization to the tricyclic tertiary carbocation **Ib**. A series of rearrangements of **Ib** and subsequent deprotonation in the C9 position to **1** accounts for 93% of the total of products. En route to the formation of **I**, mono and bicyclic triterpenes are formed, these structures include camelliol C **8**, Δ7-polypodatetraenol-9αH **11**, Δ8-polypodatetraenol **13** and Δ8(26)-polypodatetraenol-9αH **19**.

Similar rearrangement of **Ia** or **Ib** generates minor tricyclic products that account for 5.3% of total products, including the previously unknown 25-nor-9β-methylpodioda-5(10),17,21-trienol **3**. A small amount of cation **Ia** cyclizes to dammarane (0.18%) and ursane or oleanane (0.05%) triterpenes.

*Aly*PEN4 and *Ath*PEN4 are both quite product specific, with P_1/P_2 of 46 ±1 and $P_1/\sum P_i$ is 0.93 ±1. OSCs that make mainly tricycles are less frequent than OSCs making tetracyclic and pentacyclic triterpenes. The fascinating PEN4 mechanism prevents ring

expansion by forming predominantly 13αH cation **Ib**, which can form thalianol **1** but not undergo ring expansion to tetracycles and pentacycles.

The products of thalianol synthase in *A. thaliana* are further metabolized by neighboring CYP450s[22,39] (gene clusters), suggesting that thalianol oxidation products may play an important role in *Arabidopsis*. BLAST searching of the *A. thaliana* CYP450s studied by Castillo *et al.*[22] gave close matches to their *A. lyrata* orthologs, which are annotated on NCBI as CYP705A2 (XM_002870201, i.e. THAS) and CYP705A5 (XM_002863793, i.e. THAH). Thus, like LUP3 and its P450, PEN4 and its P450s have been highly conserved over the 10-million period that otherwise saw major changes in OSC phylogeny between *A. thaliana* and *A. lyrata*.

3.4.3 Mechanism and Product Profile of *A. lyrata* PEN6

A. lyrata PEN6 is 93% identical to *A. thaliana* PEN6, with two differences in active site residues. Although these two enzymes are orthologs, the major product of both enzymes, seco-β-amyrin (see Scheme 3.2), is made in different quantities (Table 3.8 and Figure 3.11). *Ath*PEN6 was previously reported by Shibuya *et al.*[23] in 2007 and a more detailed product profile is shown by Castillo *et al.*[17]

En route to formation of seco-β-amyrin **1** (72%), 28% of the substrate is lost to the formation minor products, most of which (6.3%) are oleananes. Another 7.6% are ursanes, including seco-α-amyrin **12**. The two most abundant minor products are lupeol **2** and 9αH-polypoda-7,13*E*,17*E*,21-tetraen-3β-ol **3**. Formation of **3** is not surprising since this enzyme favors deprotonation of C8 cations to make Δ7 double bonds. The rest of the material goes to the formation of other lupanes, dammaranes, tricycles and monocycles. *Aly*PEN6 makes a larger portion of non-seco byproducts (43%) than *Ath*PEN6 (25%),. Figure 3.11 compares graphically the product profile of both enzymes.

The work with PEN6 is complemented with the seed extract information shown in Table 3.9. This analysis was aided by information that Hui Shan[40] obtained from analysis of *A. thaliana* seed. She concluded that seco-β-amyrin synthase is expressed during seed formation in *A. thaliana*. Remarkably her seed triterpene profile matched closely the triterpene product profile of *Ath*PEN6 from yeast heterologous expression. My *Aly*PEN6

results and *A. lyrata* seed analyses parallel the corresponding results for *A. thaliana*. The product profile found in the seed extracts of *A. lyrata* and the product profile obtained by heterologous expression of *Aly*PEN6 are quite similar (Table 3.9). This means that the mixture created by PEN6 in both species is left mostly untouched by further metabolism and uncontaminated by triterpenes from other OSCs of secondary metabolism.

Although one can only speculate about the role of PEN6 in seeds, it seems that both organisms need seco-amyrin products in the seeds; maybe it is not about the ratio of seco-α-amyrin to seco-β-amyrin, but that the major product present in seeds has to be mostly seco-amyrin type compounds.

3.4.4 Mechanism and Product Profile of *A. lyrata* PEN9

A. lyrata PEN9 is only 78% identical to *A. thaliana* PEN6, and it lacks an ortholog in the *A. thaliana* genome. The active site residues best match the *Ath*PEN6 active site, with two differences. During oxidosqualene cyclization *Aly*PEN9 and *Ath*PEN6 form identical tetracyclic and pentacyclic cation intermediates, but by the end of the mechanistic pathway they make a very different profile of products. *Aly*PEN9 is a moderately product specific β-amyrin synthase (P_1/P_2 of 25 and $P_1/\sum P_i$ is 0.8). A detailed mechanism of product formation and product ratios are in the Results section Scheme 3.3 and Table 3.9.

The mechanism initiates a multi-step cyclization of (3*S*)-oxidosqualene to a 6-6-6-5 carbocationic intermediate with an all-chair configuration. The resulting dammarenyl C20 cation could have its side chain up (17β) or down (17α) and this change governs the stereochemistry of the derivative products (see ref. 44 for an alternative mechanistic option). Butyrospermol **9** (1.1%) originates from the 17β epimer after three consecutive 1,2-shifts and finally elimination of the axial H7α. If water is appropriately present in the active site, dammarenediols **7** and **10** are formed (1.4% and 0.9%) by hydroxylation of the C20 cation. Alternatively, a simple proton elimination yields dammara-20,24-die-3β-ol **2** (3.2%). D-ring expansion followed by E-ring closure gives the lupanyl cation **IV**. Direct deprotonation without rearrangement gives lupeol **3** in 2.7% yield. The E-ring of **IV** can undergo ring expansion to form either germanicyl cation **V** or taraxasteryl cation **VI**. Cation **V** can lose a proton from C18 without rearrangement to form germanicol **6** (1.8%).

On the other hand, a series of 1,2- shifts followed by loss of H13α generates δ-amyrin **8** (1.2%). After formation of **VI**, a series of rearrangements by 1,2-shift followed by deprotonation results in the formation of α-amyrin **4** (2.2%). Proton elimination from **VI** can yield taraxasterol **11** (0.5%) or ψ–taraxasterol **5** (2.0%). Finally, the intermediate cation **V** can follow a series of hydride shifts, with elimination of H12α giving β-amyrin **1** (81%).

A. thaliana has two β-amyrin synthases located in the LUP clade, *Ath*LUP4 and *Ath*LUP2. *Ath*LUP4 makes 74% β-amyrin and is considered to account for the formation of most of the β-amyrin found in *A. thaliana* stem wax. [41, 42, 43] *A. lyrata* does have an ortholog to *Ath*LUP2 (34% β-amyrin, sometimes called a mixed amyrin synthase); however, it lost the LUP4 ortholog. One can speculate that maybe the ancestral parent of *A. lyrata* and *A. thaliana* had two sets of β-amyrin synthases; during species divergence, each plant conserved only one pair of β-amyrin synthases, which became characteristic of each organism. Preliminary results of *A. lyrata* plant extracts shows that, as well as in *A. thaliana*, the major product of leaf wax is β-amyrin.

3.5 Final remarks

The work presented in this chapter demonstrates the importance of model organisms in the study of natural products. The comprehensive study of triterpene biosynthesis in *A. thaliana* by heterologous yeast expression was the first step to understand the complex mechanism of formation of triterpene in plants. The work done in *A. thaliana* was the guide to study triterpene biosynthesis in the closely related organism, *A. lyrata*. With the availability of both genomes and the experimental data that was obtained for *A. thaliana*, the analysis of *A. lyrata* was easier to tackle.

Genomic DNA analysis, active site comparison and protein expression were applied to the study of *A. lyrata* to determine the number of cyclases present in this organism, and the possible product profiles of each cyclase without the necessity of expression of the protein; especially in cyclases that were extremely similar to their orthologs.

Based on the DNA study it seems likely that cyclases with high sequence similarity and identical active site residues would make the same major products as their *A. thaliana*

counterpart. When sequence similarity was high, but changes in the active site existed (for example, with *Aly*PEN4 and *Aly*PEN6), expression was required in order to verify if this would affect the major product or the product profile.

Finally, cyclases that were discovered to be unique to this organism (*Aly*PEN8, *Aly*PEN9, *Aly*LUP6) were studied by DNA sequencing and/or protein expression, so that it could be determined if they were functional OSCs and what their product profile might be. When an ortholog is not available, it is hard to make any conclusion on the product profile of that particular OSC based only on sequence analysis. Protein expression has to be done in order to complete the triterpene product profile.

At the end of this study, nine OSCs were proposed to be part of *A. lyrata* and only three OSCs were required to be expressed. The analysis of this plant was much easier than the model organism *A. thaliana*, because of all the background data that was available.

REFERENCES

[1] Hu, T. T.; Pattyn, P.; Bakker E. G.; Cao. J.; Cheng J. F.; Clark' R. M.; Fahlgren, N.; Fawcett, J. A.; Grimwood, J.; Gundlach, H.; Haberer, G.; Hollister, J. D.; Ossowski, S.; Ottilar, R. P.; Salamov, A. A.; Schneeberger, K.; Spannagl, M.; Wang, X.; Yang, L.; Nasrallah, M. E.; Bergelson, J.; Carrington, J. C.; Gaut, B. S.; Schmitz, J.; Mayer, K. F. X.; Van de Peer, Y.; Grigoriev, I. V.; Nordborg, M.; Weigel, D.; Guo, Y. *Nat. Genet.* **2011**, *43*, 476-481.

[2] Goodstein, D. M.; Shu, S.; Howson, R. Neupane R.; Hayes, R. D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; Rokhsar, D. S. *Nucleic Acids Res*. **2012**, *40*, D1178-D1186.

[3] SeqMan NGen®. Version 5.03 DNASTAR. Madison, WI. http://www.dnastar.com/t-nextgen-seqman-ngen.aspx

[4] Geer L. Y.; Marchler-Bauer, A.; Geer, R. C.; Han, L.; He, J.; He, S.; Liu C.; Shi, W.; Bryant, S. H. *Nucleic Acids Res***. 2010**, *38*, D492-D496.

[5] Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. *Mol. Biol Evol*. **2013**, *30*, 2725-2729.

[6] The average OSC length is 760 aminoacids.

[7] Thoma, R.; Schulz-Gasch, T.; D'Arcy, B.; Benz, J.; Aebi, J.; Dehmlow, H.; Hennig, M.; Stihle, M.; Ruf, A. Nature **2004**, *432*, 118-122.

[8] Fazio, G. C.; Xu, R.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2004**, *126*, 5678-5679.

[9] Shan, H.; Wilson, K. W.; Castillo, D. A.; Matsuda, S. P. T. *Org. Lett.* **2015**, *17*, 3986-3989.

[10] Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Kolesnikova, M. D.; Onak, C. S.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2007**, *129*, 11213-11222.

[11] Kolesnikova, M. D.; Obermeyer, A. C.; Wilson, W. K.; Lynch, D. A.; Xiong, Q.; Matsuda, S. P. T. *Org. Lett.* **2007**, *9*, 2183-2186.

[12] Itoh, T.; Tamura, T.; Matsumoto, T. *Lipids* **1976**, *11*, 434-441.

[13] Mills, J. S. *J. Chem. Soc.* **1956**, 2196-2202.

[14] Justicia, J.; Rosales, A.; Bunuel, E.; Oller-Lopez, J.L.; Valdivia, M.; Haidour, A.; Oltra, J.E.; Barrero, A.F.; Cardenas, D.J.; Cuerva, J.M. *Chem. Eur. J.* **2004**, *10*, 1778- 1788.

[15] Goad, L.J.; Akihisa, T.; Analysis of Sterols; Blackie (Chapman & Hall): London, **1997**.

[16] Kushiro, T.; Shibuya, M.; Masuda, K.; Ebizuka, Y. *Tetrahedron Lett.* **2000**, *41*, 7705-7710.

[17] Castillo, D. A. *Ph.D. Dissertation*, Rice University **2014**.

[18] David, S. *Bull. Soc. Chim. Fr.* **1949**, 155-160. (b) David, S. *Bull. Soc. Chim. Fr.* **1950**, 169-172.

[19] Ames, T. R.; Beton, J. L.; Bowers, A.; Halsall, T. G.; Jones, E. R. H. *J. Chem. Soc.* **1954**, 1905-1919.

[20] Bischof, B.; Jeger, O.; Ruzicka, L. *Helv. Chim. Acta* **1949**, *32*, 1911-1921.

[21] Chatterjee, A.; Kundu, A. B.; Chakrabortty, T.; Chandrasekharan, S. *J. Chem. Soc. Commun.* **1968**, 418-419.

[22] Castillo, D.A.; Kolesnikova, M.D.; Matsuda, S.P.T. *J. Am. Chem. Soc.* **2013**, *135*, 5885-5894.

[23] Shibuya, M.; Xiang, T.; Katsube, Y.; Otsuka, M.; Zhang, H.; Ebizuka, Y. *J. Am. Chem. Soc.* **2007**, *129*, 1450-1455.

[24] Ames, T. R.; Halsall, T. G.; Jones, E. R. H. *J. Chem. Soc.* **1951**, 450-457.

[25] (a) Lahey, F. N.; Leeding, M. V. *Proc. Chem. Soc.* **1958**, 2196-2202. (b) Fukuoka, M.; Natori, S. *Chem. Pharm. Bull.* **1972**, *20*, 974-979.

[26] Bauer, S.; Schulte, E.; Their, H. P. *Eur. Food Res. Technol.* **2004**, *219*, 487–491.

[27] (a) Herrera, J. B. R.; Bartel, B.; Wilson, W. K.; Matsuda, S. P. T. *Phytochemistry* **1998**, *49*, 1905-1911. (b) Segura, M. J. R.; Meyer, M. M.; Matsuda, S. P. T. *Org. Lett.* **2000**, *2*, 2257-2259.

[28] (a) Khastgit, H. N.; Segupta, P. *Chem Ind.* **1961**, 1077-1078. (b) Segupta, P.; Khastgit, H. N. *Tetrahedron* **1963**, *19*, 123-132.

[29] Wang, Z.; Yeats, T.; Han, H.; Jetter, R. *J. Biol. Chem.* **2010**, *285*, 29703–29712

[30] (a) Laird, W.; Spring, F. S.; Stevenson, R. *J. Chem. Soc.* **1961**, 2638− 2642. (b) Agata, I.; Corey, E. J.; Hortmann, A. G.; Klein, J.; Proskow, S.; Ursprung, J. *J. Org. Chem.* **1965**, 30, 1698−1710. (c) Chivers, H.; Corbett, R. E.; Heng, C. K.; Wilkins, A. L. *J. Chem. Soc. C* **1973**, 1814-1816.

[31] Cao, D.; Su Y. L.; Yang J. S. *Acta Pharmacol. Sin.* **1991**, *27* (6), 445-451.

[32] Brooks, C. J. W. *Chem. Ind.* **1953**, 1178.

[33] Akihisa, T.; Yasukawa, K.; Kimura, Y.; Takase, S.-I.; Yamanouchi, S.; Tamura, T. *Chem. Pham. Bull.* **1997**, *45*, 2016-2023.

[34] Su, Y.; Tachibana, S.; Sumimoto, M. *Mokuzai Gakkaishi* **1986**, *32*, 190−202.

[35] Talapatra, S. K.; Sengupta, S.; Talapatra, B. *Tetrahedron Lett.* **1968**, 5963-5968.

[36] Ferreira, M.-J. U.; Lobo, A. M.; O'Mahoney, C. A.; Williams, D. J.; Wyler, H. *J. Chem. Soc., Perkin Trans. 1* **1990**, 185-187.

[37] Field, B.; Fiston-Lavier, A. S.; Kemen, A.; Geisler, K.; Quesneville, H.; Osbourn, A. E. *Proc. Natl. Acad. Sci.* **2011**, *108*, 16116-16121.

[38] (a) Hart, E. A.; Hua, L.; Darr, L. B.; Wilson, W. K.; Pang, J.; Matsuda, S. P. T., *J. Am. Chem. Soc.* **1999**, *121*, 9887-9888. (b) Ito, R.; Hashimoto, I.; Masukawa, Y.; Hoshino, T. *Chem. Eur. J.* **2013**, *19*, 17150−17158.

[39] Field, B.; Osbourn, A. E. *Science* **2008**, *320*, 543-547.

[40] Shan, H.; Willson, W. K. unpublished results.

[41] Shan, H.; Wilson, W. K.; Phillips, D. R.; Bartel, B.; Matsuda, S. P. T. *Org. Lett.* **2008**, *10*, 1897-1900.

[42] Buschhaus, C.; Jetter, R. *Plant Physiol.* **2012**, *160*, 1120-1129.

[43] Shibuya, M.; Katsube, Y.; Otsuka, M.; Zhang, H.; Tansakul, P.; Xiang, T.; Ebizuka, Y. *Plant Physiology and Biochemistry*, **2009**, *47*, 26-30.

[44] Morlacchi, P.; Wilson, W. K.; Xiong, Q.; Bhaduri, A.; Sttivend, D.; Kolesnikova, M. D.; Matsuda, S. P. T. *Org. Lett.* **2009**, *11*, 2627-2630.

CHAPTER 4

Exploring Triterpene Biosynthesis in the *Lactuca* Clade: Expressed-Sequence Tags (ESTs) Mining Reveals the First Nematocyphol Synthase

This chapter describes the product profile of *Lactuca perennis* nematocyphol synthase (NEM), which is encoded by the EST/cDNA name *CLPX12873* (GenBank Acc: DW077158).[1] Because the *Lactuca* family has high sequence similarity between OSCs of different species (see Chapter 5), the L. perennis ortholog was chosen to be expressed in place of the already available pGCF18.5 construct, which has a frameshift mutation that created a pseudogene and explains the lack of expression.[2] DW077158, with its almost identical active site amino acid sequence and high level of sequence similarity (94%) to Gia's construct,[2] was the closest homolog available.

Work on this cyclase began prior to the availability of the *L. sativa* genome. Under those circumstances, the only possible source of cDNA material for *Lactuca* was expressed sequence tags (ESTs). ESTs mining approaches were the only way to find cyclases that are in the Asteraceae family,[2] although this strategy only provides an incomplete picture of triterpene metabolism in *Lactuca*.

Products are numbered from highest to lowest amount, starting with **1** for the major product (Table 4.2). Gene sequences, spectral data, and additional information can be found in the Appendix.

4.1 Background and Previous Work

The *Lactuca* genus is part of a large family of angiosperms called Asteraceae. The Asteraceae or Compositae family is a sister group to the Eurosids of which the reference plant, *Arabidopsis thaliana* is a member.

As it has been seen with the model organism *Arabidopsis thaliana*, availability of the genome has allowed for the generation of a comprehensive accounting of triterpenes in this plant.[3] However when the genome is not available other strategies, particularly ESTs libraries, are of enormous assistance in the process of the discovery of putative OSCs.[4]

When this project began, no *Lactuca* genomes were available, but a number of ESTs from both *L. serriola* and *L. sativa* and several other species of lettuce had been accessible for a number of years.[5] The idea of a systematic analysis of the available ESTs to find putative OSCs came from the fact that triterpene metabolites have been known to confer resistance to herbivory in other plant systems.[7] Moreover, it has been reported that traditional organic extraction from commercially available lettuce reveals several triterpene alcohols: α- amyrin, β-amyrin, germanicol and ψ-taraxasterol.[8]

One purpose of the Compositae Genome Project was to study the differences between genomes so they could find the genes that are involved in pathogen resistance, since both species exhibit genetic variance at disease resistance loci.[6] This could potentially facilitate engineering desired traits into lettuce, which is an agriculturally important crop.

4.1.1 Identification of a Putative OSC via Basic Local Alignment Search Tool

The Arizona Genomic Institute made publicly accessible EST libraries of a number of Lactuca species available at an affordable price. Using *Arabidopsis thaliana* cycloartenol synthase and basic local alignment search tool (BLAST), Gia Fazio[2] identified several putative OSCs from *L. sativa* and *L. serriola*. Among them, she found one encoded by the *L. serriola* clone BU003176/QGG34F12 that was subcloned into an expression vector. The final construct was named pGCF18.5; however this construct contained a frameshift mutation. This mutation was determined to be a PCR error and, as described in Chapter 5, was fixed using site directed mutagenesis. However, neither the EHY41 nor RXY6 yeast transformants gave any triterpene product. At this point I abandoned this strategy and decided to BLAST BU003176 to search for a close homolog among the pool of other Lactuca ESTs. Among promising candidates, I chose the DW077158 sequence from *L. perennis*, from which pMMG12.3 was constructed and successfully expressed in RXY6 and EHY41, as described below. The open question remains whether the *L. serriola* ortholog is a pseudogene or contained a lab construction flaw.

Using BU003176 sequence from *L. serriola*, a protein query versus translated sequence (tBLASTn) search was performed against the "ESTs Asterids" database. This BLAST search resulted in three good hits from *L. perennis* that covered parts of the

BU003176 sequence; those ESTs are: DW094964, DW076889 and DW077158. SeqMan was used to assemble these three sequences using BU003176 as a template. As it turns out, these three ESTs belong to the same OSC,but cover different regions. The DW077158 clone was chosen to be expressed because it was already available in our plasmid library. Additional ESTs, *L. virosa* (DW153103) and *L. sativa* (DY975354 and DY975590), were obtained from the BLAST search; this suggests that BU003176 could have close homologs in other species of *Lactuca*.

The next section will provide details about phylogeny, plasmid and strain construction, along with purification and characterization of nematocyphol synthase (NEM) products.

4.2 Experimental Procedures

4.2.1 Plasmid and Yeast Strain Construction

4.2.1.1 Cloning of Nematocyphol Synthase

The clone DW017758 was obtained from Arizona Genome Institute AGI (www.genome.arizona.edu). Sequencing with pertinent primers confirmed the ~2.3 kbp cDNA sequence. *SalI* and *NotI* restriction sites were added by PCR amplification, and the amplicon was subcloned into pGEM-T vector system. The construct was then digested with *SalI* and *NotI* restriction enzymes and gel purified, and cloned into the yeast expression vectors pRS424Gal and pRS426Gal carrying the *TRIP1* and *URA3* selection markers, respectively. The resultant plasmids containing the ~2.3 kbp gene were named pMMG12.2 and pMMG12.3, respectively. The final constructs were sequenced with T3 and 426GalF primers. Sequencing confirmed that the pMMG12.2 and pMMG12.3 each contained a 2.3 kpb coding sequence equivalent to DW017758.

4.2.1.2 Yeast Strains Construction

The plasmid pMMG12.3 was used to transform *Saccharomyces cerevisiae* strain RXY6 (*ERG1* and *ERG7* mutant) and the plasmid pMMG12.2 was used to transform *Saccharomyces cerevisiae* strain EHY41 (containing a single-point mutation in the UPC2 transcription factor (*upc2-1*) and overexpression of trHMGR1) using the lithium acetate method. RXY6 transformants were selected on synthetic complete medium lacking uracil, solidified with 1.5% agar, and supplemented with 2% glucose, 13 mg/L hemin chloride, 20 mg/L ergosterol, and 5 g/L Tween 80. EHY41 transformants were selected on synthetic complete medium lacking tryptophan, solidified with 1.5% agar, and supplemented with 2% glucose.

4.2.2 *In vivo* Experiments toward Characterizing the Nematocyphol Synthase Product Profile

An 8-L culture of EHY41[pMMG12.2] was grown in synthetic complete medium lacking tryptophan and containing 2% galactose as a carbon source. The cultures were grown to saturation at 30 ˚C with shaking at 250 rpm. After centrifugation, the 174 g pellet was saponified with 5 mL of 10% KOH (w/v) in 80% EtOH (v/v) per gram of cells for 4 h at 70 ˚C. Following saponification, the supernatant was separated by decantation, and the excess ethanol was removed by evaporation. The aqueous layer was then extracted with 8 x 100 mL hexanes. The combined hexanes layers were concentrated to dryness by rotatory evaporation in a round-bottomed flask, before being transferred to a pre-weighed scintillation vial and evaporated under nitrogen. The crude NSLs (325 mg) and were analyzed by GC-MS (Figure 4.1) and 30 mg was used for NMR.

The crude NSL were dissolved in methylene chloride and loaded onto a 10 g SiO_2 gel solid phase extraction column (Luknova). Thirty-seven 8 mL fractions were collected (via an SPE vacuum manifold) with methylene chloride as eluent, after which 8 additional fractions were collected with diethyl ether (Et_2O). Each fraction was then analyzed by GC-MS and those containing triterpene alcohols (fractions 4-13) and triterpene diols (fractions 31-33) were separately combined.

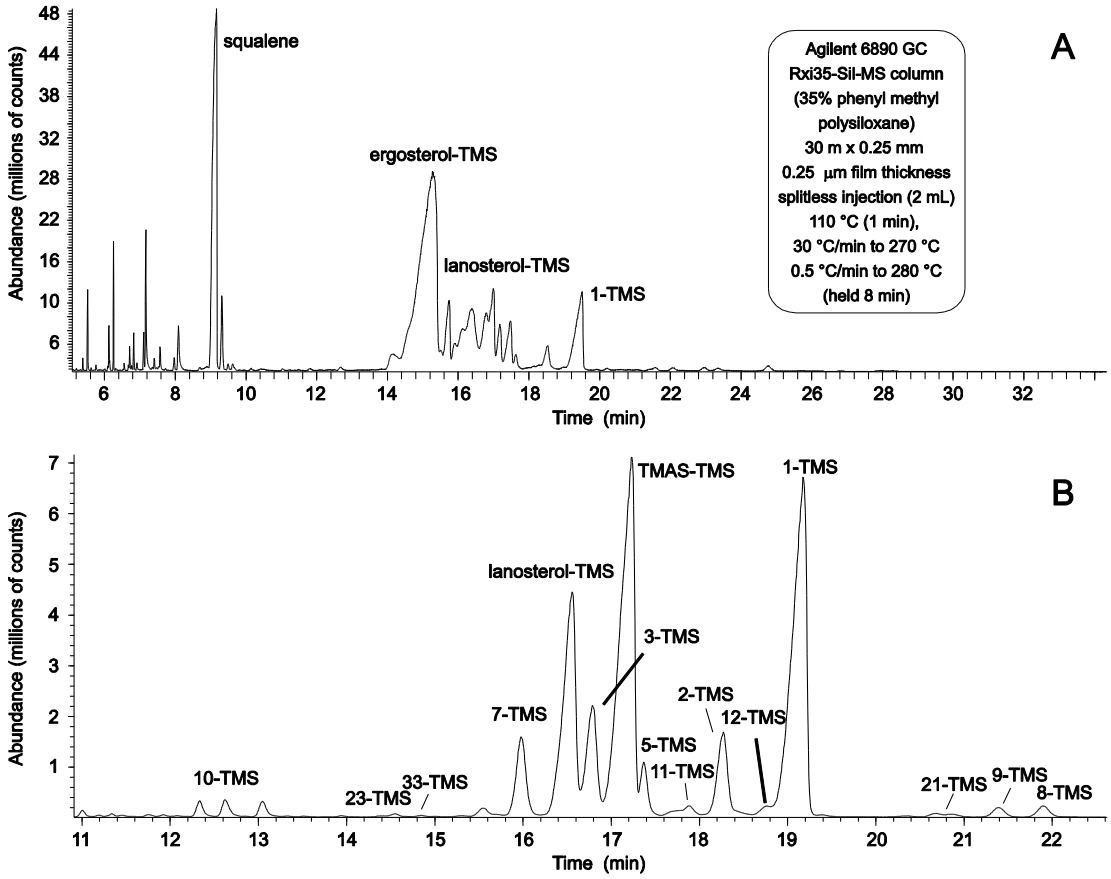


Figure 4.1. a) GC-MS chromatograms from the 8 L EHY41[pMMG12.2] experiment: a) crude NSL extract and b) combined fractions of the SPE purified triterpenes.

Additional purification of triterpene (alcohol) fractions was necessary for further separation of minor products. In addition, triterpene diol fractions also needed further purification since ergosterol was still a major component and would lower the quality of NMR spectra. The combined triterpene fractions (fractions 4-13 and 31-33) were thus dissolved in 2 mL of hexanes and loaded onto a 5 g SiO_2 gel SPE column (Phenomenex); 60 x 8 mL fractions were collected using diethyl ether (Et_2O)/ hexanes gradients (0%, 1%, 2%, 3%, 5%, 10%, 20%, 30%, 50%, and 100% diethyl ether). Fractions containing the triterpene alcohols and triterpene diols were combined to yield ca. 64 mg of triterpene

alcohols and ca. 5 mg of triterpene diols. Half of the combined fractions was stored and the other half was used for GC-MS and NMR analysis (^{1}H NMR and HSQC).

After NMR analysis the samples were recovered from the NMR tube by evaporation into a 1 dram vial. The triterpene alcohol sample (34 mg) was then prepared for preparative reverse phase HPLC separation/purification. The sample was dissolved in 300 µL of methylene chloride, followed by addition of 700 µL of MeOH. The mixture was filtered through a glass wool plug to remove any insoluble particles out of the sample. Following equilibration of the HPLC system, the 1 mL sample was injected onto a Phenomenex Prodigy C_{18} column (250 x 21.2 mm, 5 µm particle size), and eluted with a linear gradient of 90-100% MeOH and a flow rate of 8 mL/min, UV detection at 210 nm, and collection of 1 min fractions. Fractions 30 to 80 were analyzed by GC-MS.

Fractions 59, 60, 61 64 and 65 contained unfamiliar triterpenes. Fraction 59 contained 17-epilupeol (**24**); fractions 60 and 61 neolupenol (**7**), tarolupenol (**6**) and DB-friedolup-5-en-3β-ol (**19**); fraction 64 tylolupenol A (**3**) and lup-19(21)-en-3β-ol (**13**) and fraction 66 and 67 nematocyphol (**1**). Fractions 59, 60, 61, 64 and 65 were analyzed by GC-MS, ^{1}H NMR, HSQC and in some cases COSYDEC, HMBC and DEPT.

Fractions 59, 60-61, 64, and 65 were further separated by analytical HPLC. Samples were dissolved in 200 µL of methylene chloride, followed by addition of 800 µL of MeOH, and filtered through a glass wool plug prior to injection to remove any solid particles. The samples were injected onto an Imtakt Cadenza C_{18} column (250 x 4.6 mm, 3 µm particle size) and eluted with a linear gradient from 88% to 100% MeOH, with a flow of 1 mL/min, UV detection at 210 nm, and collection of 1 min fractions. For each HPLC separation, fractions 40 through 60 were taken for analysis on GC-MS and NMR.

4.2.3 *In vitro* Experiments toward Characterizing the Nematocyphol Synthase Product Profile

A cell pellet (40 g) was obtained from a 2.5 L culture of RXY6[pMMG12.3] grown at 30 ˚C, 250 rpm in synthetic complete medium lacking uracil, with 2% glucose as a carbon source, supplemented with ergosterol (20 µg/mL) and hemin chloride (13 µg/mL), which was harvested when an OD_{600} of approximately 6.0 was reached. The cell pellet

was resuspended in 0.1 M sodium phosphate buffer (40 mL, pH 7) and then racemic oxidosqualene (OS) solubilized in Triton X-100 was added to give a final concentration of 1 mg of racemic OS per mL of homogenate. The homogenate was divided into 46 microcentrifuge tubes containing ~100 borosilicate glass beads. The cells were lysed using a cell disrupter 3 x 3 min with 5 min intervals on ice. After 24 h incubation at room temperature (ca. 23 °C), the reactions were quenched with two volumes of ethanol, and the cell debris was removed by centrifugation. Following centrifugation the ethanol was almost completely removed by rotatory evaporation. The remaining aqueous phase was partitioned between hexanes and water and extracted with 6 x 100 mL hexanes. The hexanes layers were combined, washed with 50 mL of brine and dried over anhydrous Na_2SO_4 to give ca. 250 mg of crude extract, which was analyzed by GC-MS.

The sample was then purified by SPE (solid phase extraction). It was dissolved in methylene chloride and loaded onto a 10 g SiO_2 gel column (Luknova). Thirty-seven 8 mL fractions were collected (via an SPE vacuum manifold) with methylene chloride as eluent, after which 8 additional fractions were collected with diethyl ether (Et_2O). Each fraction was then analyzed by GC-MS, and those containing triterpene alcohols (fractions 4-10) and triterpene diols (fractions 30-35) were combined.

Triterpene diol fractions were collected and separated for GC-MS and NMR analysis. However, additional purification of triterpene fractions was necessary for further removal of the traces of ergosterol and oxidosqualene. The combined triterpene fractions (4-10) were thus dissolved in 2 mL of hexanes, and loaded onto a 5 g SiO_2 gel SPE column (Phenomenex); 60 x 8 mL fractions were collected using the diethyl ether (Et_2O)/ hexanes gradient described in Table 4.1

Fractions containing the triterpene alcohols and triterpene diols were combined to yield 3.5 mg of triterpene alcohols and ~1 mg of triterpene diols. These samples were used for GC-MS and NMR analysis. Proton NMR and HSQC spectra were collected for this sample.

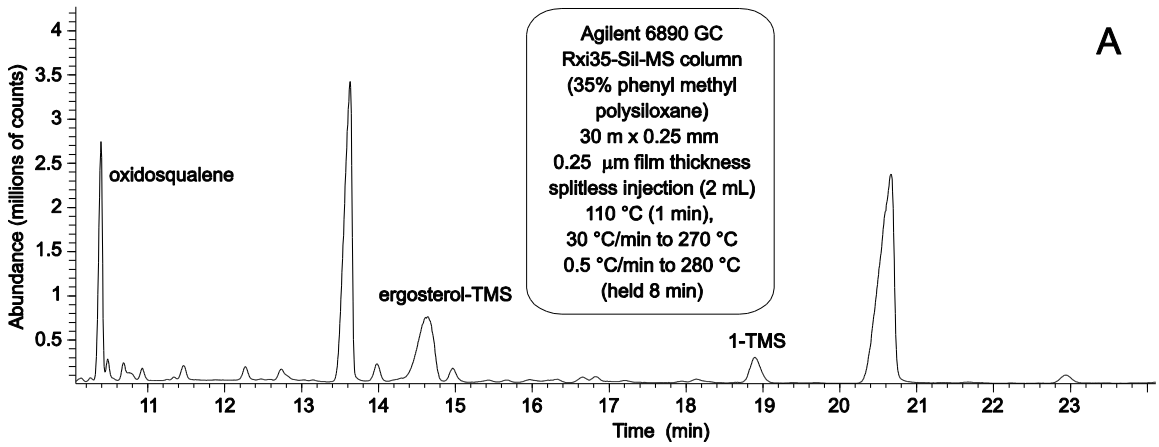


Figure 4.2. a) GC-MS chromatogram of the crude NSLs extract aliquot of the 2.5 L RXY6[pMMG12.3] experiment.

Table 4.1. SPE solvent gradient for purification of RXY6[pMMG12.3] products.

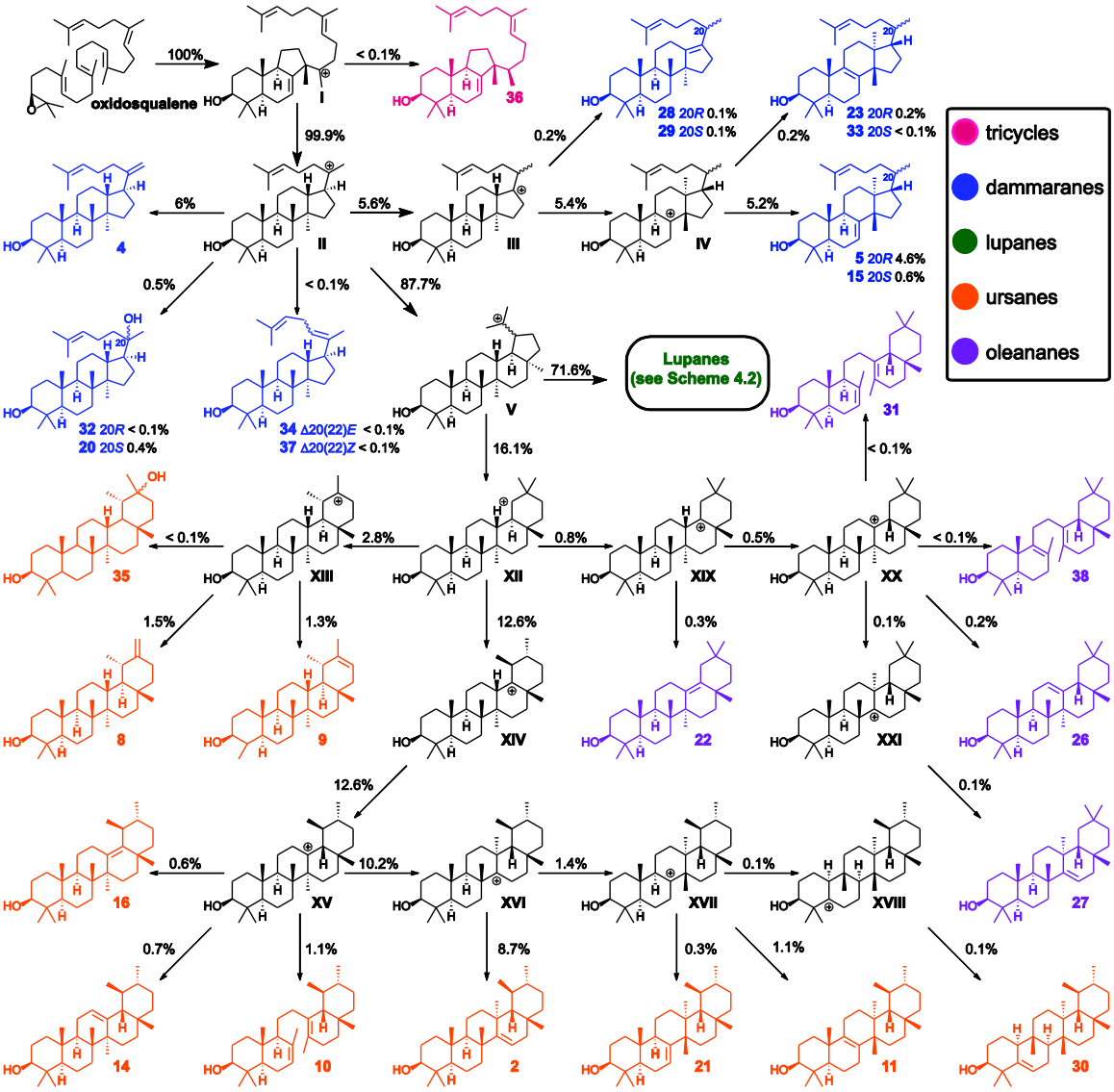
Fraction	% Et_2O in Hexanes	Fraction	% Et_2O in Hexanes
1-2	0	27-32	10%
3-6	1%	33-38	20%
7-12	2%	39-46	30%
13-18	3%	47-54	50%
19-26	5%	55-60	100%

4.3 Results

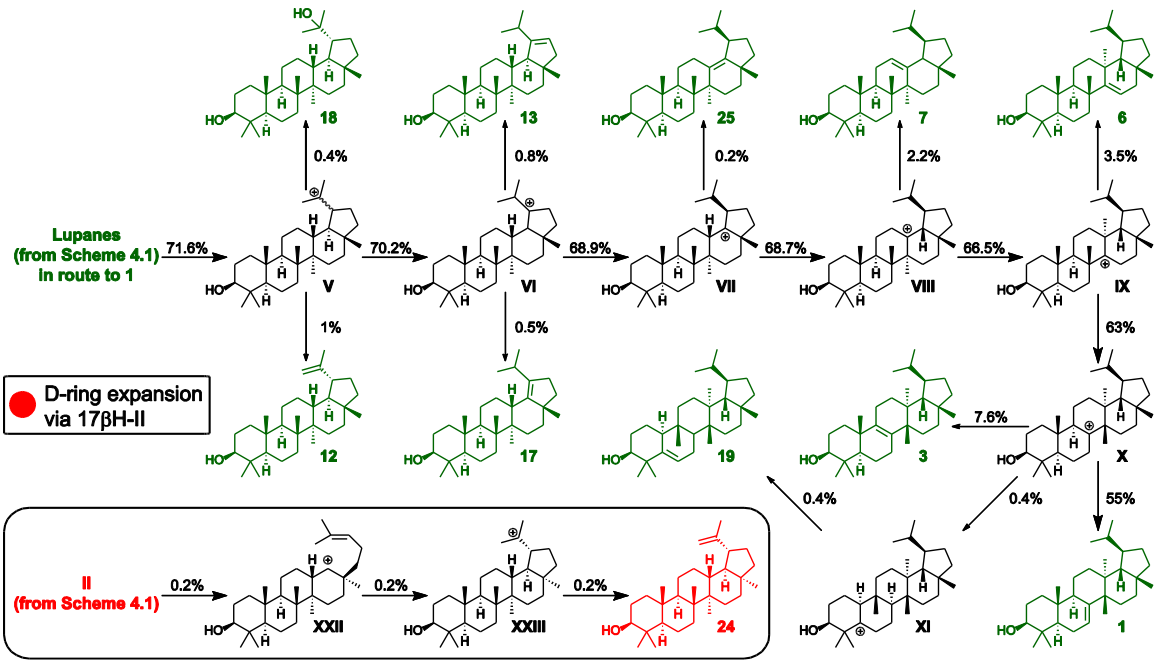
4.3.1 *In vivo* and *in vitro* Product Profile of *Lactuca perennis* Nematocyphol Synthase

Nematocyphol synthase (NEM) was named after its major product nematocyphol (**1**), which accounts for 55% of its total products. In addition to nematocyphol (**1**), this cyclase makes at least 37 additional triterpenes, more products than reported for any other OSC (but fewer than the 40 products I observed in Chapter 3 for AlyPEN6).

The enzyme mechanism and product profile are presented in Scheme 4.1 and Scheme 4.2, as well as Figure 4.3, which shows a graphical comparison of the product profile ratios for the *in vivo* and *in vitro* experiments. Product profiles of triterpene alcohols and triterpene diols were determined mainly by HSQC mixture analysis.[9] Ratios for the diols **20**, **32** and **35** were determined from the HSQC spectrum of the NSL, whereas the ratios for the triterpene alcohols were determined using the HSQC spectrum of EHY41[pMMG12.2] fractions18-34 GC-MS analysis of HPLC fractions confirmed the presence of many minor products. Table 4.2 presents a complete picture of the spectroscopic evidence for each compound of the product profile.



Scheme 4.1. Proposed mechanism of product formation by nematocyphol synthase. Formation of products from rearrangement of cation **V** to **1** and other rearranged lupeols is shown in Scheme 4.2.



Scheme 4.2. Mechanism of product formation en route to and beyond nematocyphol (**1**).

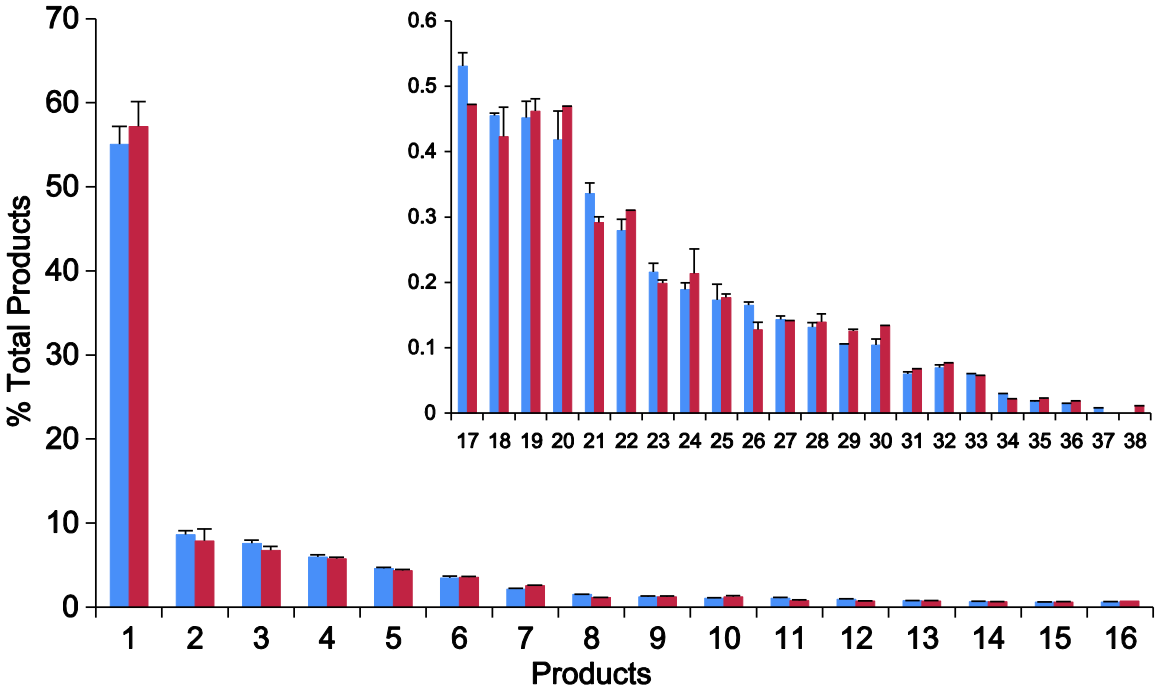


Figure 4.3. Product profile of NEM determined from an 8 L *in vivo* experiment (blue bars) and a 4 L *in vitro* experiment (red bars). NMR error bars correspond to standard deviations in NMR signal areas.

Table 4.2. Summary of analytical evidence for *L. perennis* nematocyphol synthase products identified in the 8 L *in vivo* experiment.

Product #	Product Name	% of Total	GC-MS*	NMR Evidence	Ref
1	nematocyphol	55	A,B	HSQC, 1D	20
2	isoursenol	8.7	A,B	HSQC, 1D	21
3	tylolupenol A	7.6	A,B	HSQC, 1D	22
4	dammara-20,24-dien-3β-ol	6.0	A,B	HSQC, 1D	23
5	butyrospermol	4.6	A,B	HSQC, 1D	23
6	tarolupenol	3.5	A,B	HSQC, 1D	24
7	neolupenol	2.2	A,B	HSQC, 1D	25
8	taraxasterol	1.5	A,B	HSQC, 1D	26
9	ψ-taraxasterol	1.3	A,B	HSQC, 1D	26
10	seco-α-amyrin	1.1	A,B	HSQC, 1D	27
11	isobauerenol	1.1	A,B	HSQC, 1D	28
12	lupeol	1.0	A,B	HSQC, 1D	29
13	lup-19(21)-en-3β-ol	0.8	A,B	HSQC, 1D	none
14	α-amyrin	0.7	A,B	HSQC, 1D	30
15	tirucalla-7,24-dien-3β-ol	0.6	A,B	HSQC, 1D	31
16	γ-amyrin	0.6	B	HSQC, 1D	32
17	18-lupane-3β-ol	0.5	A,B	HSQC, 1D	33
18	lupane-3β,20-diol	0.4	A,B	HSQC, 1D	34
19	DB-friedo-lup-5-en-3β-ol	0.4	A,B	HSQC, 1D	35
20	(20*S*)-dammar-24-ene-3β,20-diol	0.4	A,B	HSQC, 1D	36
21	bauerenol	0.3	A,B	HSQC, 1D	37
22	δ-amyrin	0.3	A	HSQC, 1D	38
23	euphol	0.2	A,B	HSQC, 1D	39
24	17-epilupeol	0.2	A,B	HSQC, 1D	9
25	neolup-13(18)-enol	0.2	A,B	HSQC, 1D	40
26	β-amyrin	0.2	A,B	HSQC, 1D	41
27	taraxerol	0.1	A,B	HSQC, 1D	42
28	isoeuphol	0.1	A,B	HSQC, 1D	43
29	isotirucallol	0.1	A,B	HSQC, 1D	43
30	rhoiptenol	0.1	A,B	HSQC, 1D	44
31	seco-β-amyrin	<0.1	A,B	HSQC, 1D	27

32	(20*R*)-dammar-24-ene-3β,20-diol	<0.1	A,B	HSQC, 1D	36
33	tirucalla-8,24-dien-3β-ol	<0.1	A,B	HSQC, 1D	45
34	dammara-20(22)*E*,24-dien-3β-ol	<0.1	A,B	HSQC, 1D	46
35	(20*R*/*S*) -taraxasteindiol	<0.1	A,B	HSQC, 1D	47
36	Δ7-thalianol	<0.1	A	HSQC, 1D	14
37	dammara-20(22)*Z*,24-dien-3β-ol	<0.1	A,B	HSQC, 1D	46
38	Δ8(26)-seco-β-amyrin	<0.1	A,B	HSQC, 1D	27

* GC-MS evidence using Method A and B, Chapter 2.
-- Not reported, compound found by Maria Kolesnikova.

One of the *L. perennis* nematocyphol synthase products, lup-19(21)-en-3β-ol (**13**), is a previously unknown compound whose structure was determined during course of this work. This compound was partially isolated from the 8 L EHY41 triterpene mixture by reversed-phase HPLC as a 3:1:1 ratio of tylolupenol A (**3**), isoursenol (**2**) and lup-19(21)-en-3β-ol (**13**). This mixture was analyzed by GC-MS, ^{1}H NMR, HSQC, HMBC, and COSYDEC. The NMR results were used to determine the structure of **13** from C-C and C-H connectivities (Figure 4.4). Quantum mechanical calculations (performed by William K. Wilson) confirmed the chemical shift assignments and the position of the double bond. This will be described in the next section.

4.3.2 Quantum Mechanical Confirmation of C-19 Stereochemistry for lup-19(21)-en-3β-ol

Quantum mechanical calculations, including GIAO predictions of NMR chemical shifts, were performed as described by Castillo *et al.*[19] The C19-C20 rotamer populations for **13** were [A 0.499, B 0.498, and C 0.003] (Figure 4.5). Comparison of observed and calculated chemical shifts in Table 4.3 and Table 4.4 validated the NMR signal assignments and the structure of **13**.

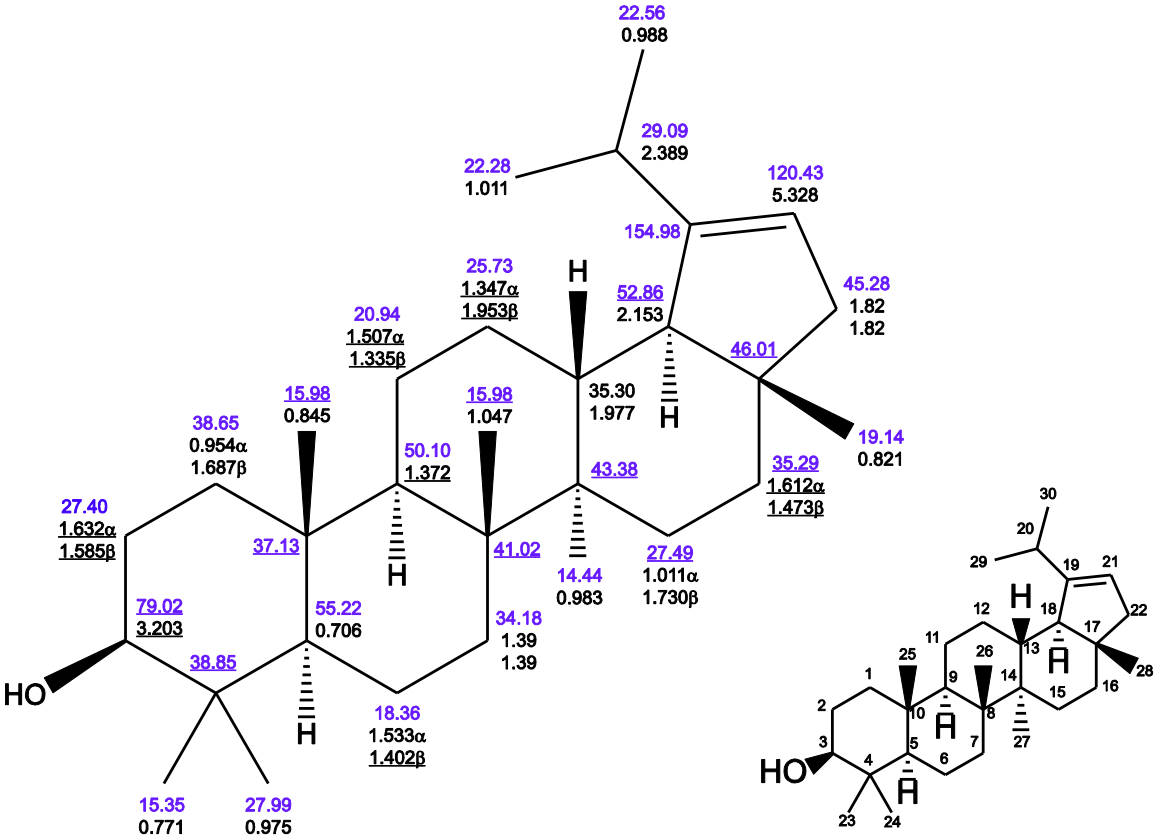


Figure 4.4. NMR assignments and atom numbering for lup-19(21)-en-3β-ol. Chemical shifts are estimated as accurate to ±0.001 ppm for ^{1}H and ±0.01 ppm for ^{13}C except for underlined values (±0.003 ppm and ±0.03 ppm) and ^{1}H values given to two decimal places (±0.01 ppm). Chemical shifts are not corrected for strong coupling.

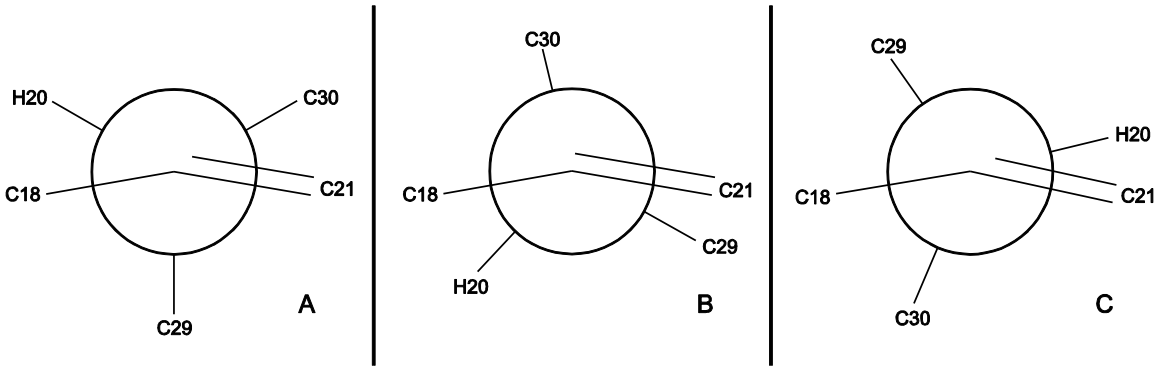


Figure 4.5 The side C19-C20 rotamers of lup-19(21)-en-3β-ol (**13**). GIAO predictions were made for each of these rotamers, whereas general chemical shift corrections for conformational heterogeneity of the C3-O rotamers were done as described previously.[19]

Table 4.3. Comparison of the observed and calculated ^{13}C chemical shifts for **13**.

^{13}C atom	Obs δ	Calc δ	Δ Calc δ - Obs δ
C-1	38.7	38.1	-0.57
C-2	27.4	28.1	0.71
C-3	79.0	78.0	-1.06
C-4	38.9	38.5	-0.36
C-5	55.2	55.1	-0.11
C-6	18.4	18.6	0.28
C-7	34.2	33.0	-1.14
C-8	41.0	41.3	0.29
C-9	50.1	50.4	0.29
C-10	37.1	36.8	-0.38
C-11	20.9	20.9	-0.03
C-12	25.7	25.9	0.21
C-13	35.3	35.3	-0.01
C-14	43.4	43.9	0.49
C-15	27.5	26.9	-0.59
C-16	34.9	34.1	-0.80
C-17	46.0	45.6	-0.42
C-18	52.9	52.5	-0.36
C-19	155.0	154.8	-0.22
C-20	29.1	29.6	0.54
C-21	120.4	121.9	1.46

C-22	45.3	45.4	0.12
C-23	28.0	27.3	-0.66
C-24	15.4	15.1	-0.23
C-25	16.0	16.2	0.25
C-26	16.0	14.8	-1.20
C-27	14.4	12.6	-1.82
C-28	19.1	19.0	-0.14
C-29	22.6	22.7	0.16
C-30	22.3	22.8	0.50
		average deviation	-0.16
		rms deviation	0.67

Table 4.4. Comparison of the observed and calculated ^{1}H chemical shifts for **13**.

^{1}H atom	**Obs δ**	**Calc δ**	**Δ Calc δ - Obs δ**
H-1a	0.95	0.93	-0.02
H-2b	1.69	1.67	-0.02
H-2a	1.63	1.62	-0.02
H-2b	1.59	1.54	-0.04
H-3a	3.20	3.18	-0.02
H-5a	0.71	0.76	0.05
H-6a	1.53	1.47	-0.06
H-6b	1.40	1.41	0.01
H-7a	1.39	1.41	0.02
H-7b	1.39	1.30	-0.09
H-9a	1.37	1.51	0.14
H-11a	1.51	1.48	-0.03
H-11b	1.34	1.41	0.07
H-12a	1.35	1.47	0.12
H-12b	1.95	1.96	0.01
H-13a	1.98	2.08	0.11
H-15a	1.01	0.91	-0.10
H-15b	1.73	1.74	0.01
H-16a	1.61	1.62	0.01
H-16b	1.47	1.45	-0.02
H-18	2.15	2.38	0.23
H-20	2.39	2.45	0.06
H-21	5.33	5.35	0.02
H-22a	1.82	1.92	0.10

H-22b	1.82	1.85	0.03
H-23	0.98	0.99	0.01
H-24	0.77	0.74	-0.03
H-25	0.85	0.78	-0.07
H-26	1.05	0.99	-0.06
H-27	0.98	0.97	-0.01
H-28	0.82	0.85	0.03
H-29	0.99	0.99	0.00
H-30	1.01	1.00	-0.01
		average deviation	0.01
		rms deviation	0.069

4.4 Discussion

4.4.1 Phylogenetic Relationship between Nematocyphol Synthase and other OSCs from Asterids

Relative to various characterized asterid OSCs, *L. perennis* nematocyphol synthase exhibits the highest amino acid sequence identity to the mixed amyrin synthases and the dammarenediol synthases (72%) as shown in Table 4.5. The phylogenetic tree in Figure 4.6 suggests that these cyclases are closer to nematocyphol synthase than the lupeol synthases, which show a lower sequence identity. Moreover, the lupeol synthases show several distinctive active site residues that are not found in NEM, the mixed amyrin synthases, and dammarenediol synthases, as shown in the active site alignment in Figure 4.7.

Although this limited analysis suggests distant phylogeny between NEM and the lupeol synthases, both groups make 6-6-6-6-5 products, unlike the other OSCs in Table 4.5. The following mechanistic analysis sheds light on these seemingly paradoxical observations.

Table 4.5. Comparison of various characterized asterid OSCs with nematocyphol synthase.

AA Sequence Identity to NEM	OSC	GeneBank ID	Family	Species	Major Product	Reference
57%	OEW	AB025343	Asterids	*Olea europea*	Lupeol	4
57%	TRW	BAA86932	Asterids	*Taraxacum officinale*	Lupeol	4
57%	BAS	EU330197	Asterids	*Artemisia annua*	β-amyrin	49
59%	PNY1	AB009030	Asterids	*Panax ginseng*	β-amyrin	50
58%	PNY2	AB014057	Asterids	*Panax ginseng*	β-amyrin	51
72%	CrAS	AFJ19235	Asterids	*Catharanthus roseus*	Mixed amyrin	52
72%	OEA	AB291240	Asterids	*Olea europaea*	Mixed amyrin	53
72%	PNA	AB265170 AEO27862	Asterids	*Panax ginseng*	Dammarenediol	54
72%	PnDS	AGS79229	Asterids	*Panax notoginseng*	Dammarenediol	55
72%	PqDS	AED99864	Asterids	*Panax quinquefolius*	Dammarenediol	56
68%	SHS1	AB609123	Asterids	*Aster tataricus*	Shionone	57
---	NEM	DW077158	Asterids	*Lactuca perennis*	Nematocyphol	---

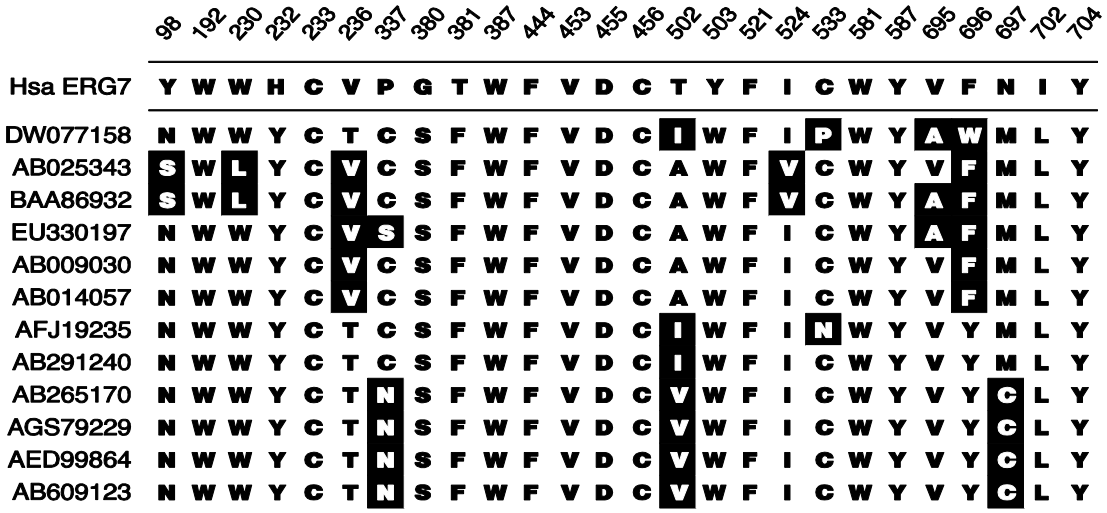
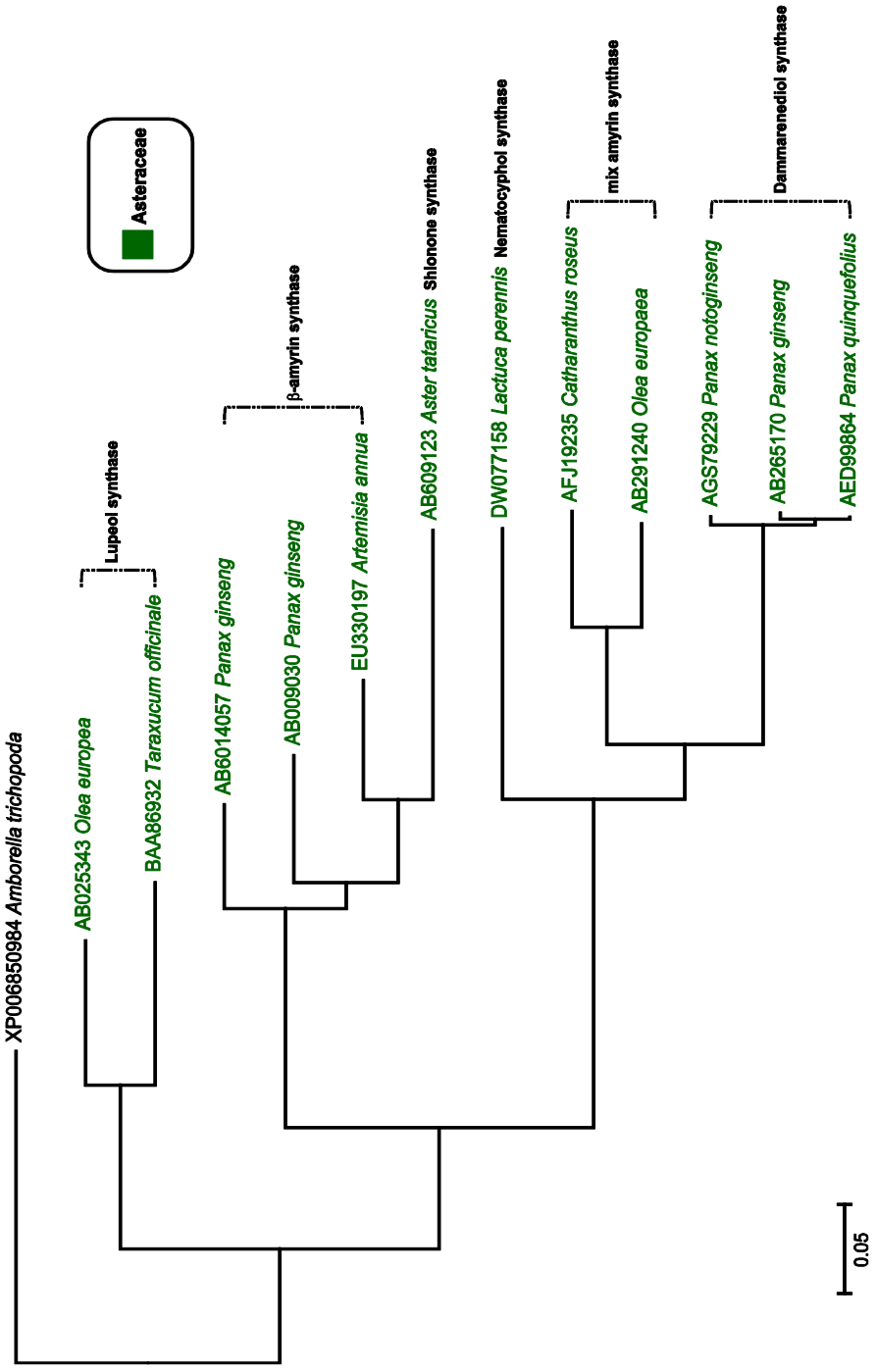


Figure 4.6. Active-site alignments of *L. perennis* nematocyphol synthase (DW0077158) and related plant OSCs from the Asterids family.



Figures 4.7. Phylogenetic tree of characterized oxidosqualene cyclases from the Asterids clade. The tree was constructed by maximum likelihood using MEGA version 6. Multiple sequence alignments were generated in MegAlign from amino acid sequences with Clustal W using default parameters.

4.4.2 Products of *L. perennis* Nematocyphol Synthase (NEM) and their Mechanism of Formation

The product profile of nematocyphol synthase was obtained from the combination of *in vivo* and *in vitro* experiments involving several purification techniques and a variety of analytical methods. Due to the high number of triterpene products that this cyclase can make, peak overlap was the major impediment in quantifying the many nematocyphol synthase products. However, HSQC analysis of EHY41 or RYX6 products enabled identification of minor NEM metabolites at a level of ~0.05% of total products due to the signal distribution along two axes.[58,60] The HSQC experiments demonstrated that besides nematocyphol (**1**), this cyclase makes at least 37 other products that account for 45% of the product profile.

The mechanism of formation and product profile are detailed in Scheme 4.1, Scheme 4.2 and Table 4.2. As shown in Scheme 4.1, 87.5% of the material goes through cation **V**, either in the 17αH intermediate **Va** or its 17βH epimer **Vb**. As suggested below, **Va** retains the 6-6-6-6-5 pentacyclic structure en route to lupane type rearranged products (71.6%), while **Vb** undergoes E-ring expansion toward the formation of ursanes and oleananes (total 16%), all of them made in less than 3% yield, except for isoursenol (**2**, 8.7%). Ursanes and oleananes (16%) are formed by E-ring expansion of **V** by C18 migration to form the germanicyl cation **XII** or by C18 migration to form a secondary cation en route to the C18 ursanyl cation **XIV**.

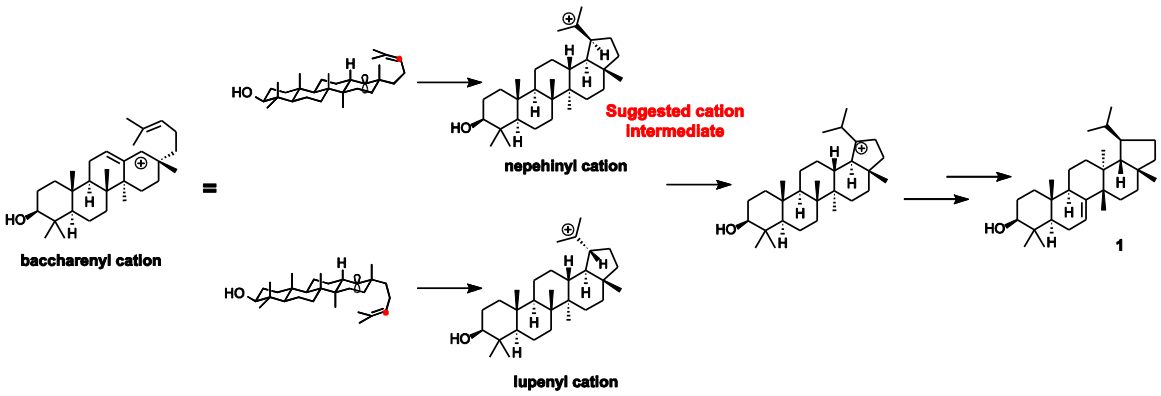
Analysis of the mechanism in Schemes 4.1 and 4.2 implies that en route to the formation of nematocyphol ~30% is lost to the formation of dammaranes, lupanes, oleananes and ursanes. Of this ~30%, 12% is dedicated to tetracycles and the rest undergoes formation of ursanes with a minor loss to oleananes.

Formation of nematocyphol **1** (55%) and a variety of lupeol derivatives is believed to arise from rearrangement and deprotonation of nepehinyl cation **Va** and account for the remaining 71.6% of material. Direct deprotonation without rearrangement of **Vb** provides lupeol **12** (1%), while water addition to C20 forms lupanediol **18** (0.4%). After a series of 1,2-shifts and deprotonation at different positions this enzyme makes 18-lup-18-en-3β-ol **17** (0.5%), neolup-13(18)-enol **25** (0.2%), DB-friedo-lup-5-en-3β-ol **19** (0.4%), neolupenol

7 (2.2%), tarolupenol **6** (3.5%), tylolupenol A **3** (7.6%), nematocyphol **1** (55%) and lup-19(21)-en-3β-ol **13** (0.8%), a previously unknown product. Once the rearrangement gets to the carbocationic intermediates **Va** and **Vb**, a series of rearrangements to **X** will cause the loss of 16.6% of the material en route to the formation of **1**.

Determination of the product profile of a cyclase allows it to be compared in accuracy with other OSCs by the standard measures P_1/P_2 and $P_1/\sum P_i$.[14] Nematocyphol synthase has a value of P_1/P_2 of 6.35 and $P_1/\sum P_i$ is 0.5. The analysis of the product profile of nematocyphol synthase and the values of P_1, P_2 and P_i shows that this enzyme can be placed in the range of moderately product-specific OSCs, as described by Lodeiro *et al.*[16]

NEM is capable of performing the series of 1,2-shifts necessary to form nematocyphol and 9 other products, and also is able to undergo ring expansion to make ursanes and oleananes. Cationic intermediate **V** is shown in Scheme 4.2 with an ambiguous stereochemistry at the C19 position. Evidence from molecular mechanics calculations and organic synthesis suggest that the lupanyl cation is not the only possible intermediate that could eventually lead to the formation of the lupane-type rearranged products.

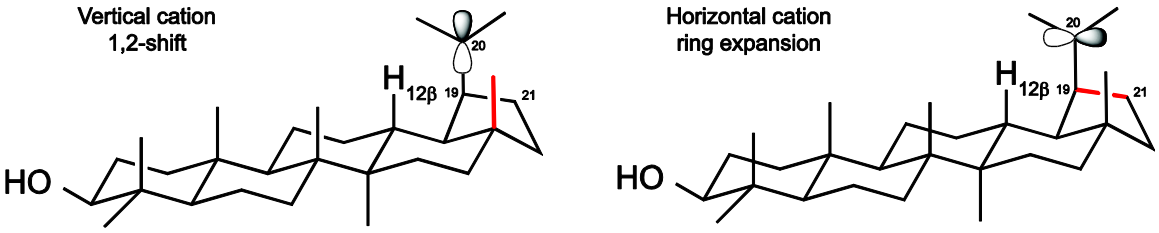


Scheme 4.3. Suggested mechanism for the formation of nematocyphol and rearranged lupane type compounds.

Hui Shan, Matias Kinzurik and Dorianne A. Castillo[60] performed a series of triflic acid isomerizations using as starting material either normal lupane-3β,20-diol (17βH) or its C17 epimer 17αH-lupane-3β,20-diol, which is here referred to as nepehine-3β,20-diol..When normal lupane-3β,20-diol was used, the product profile of this reaction was

primarily comprised of ring expansion products and rearranged lupenols were almost undetectable (<<1%). In contrast, when nepehine-3β,20-diol was used, they observed mainly rearranged lupenols, including several NEM products and ~30% ursanes and oleananes.

Preliminary results from QM calculations performed by William K. Wilson present a possible explanation for the lupanyl cation to perform ring expansion and nepehinyl cation to do rearrangements. This hypothesis involves the understanding of the reactivity of horizontal and vertical cations (Scheme 4.4).



Scheme 4.4. Illustration of vertical and horizontal cations. Thick bonds in red denote hyperconjugation with the cation. The relevant hyperconjugation for the vertical cation is with the C19-H19 bond, but the C17-C28 bond is shown in red for convenience.

In vertical carbocations, orientation of the 2p orbital leads to hyperconjugation with axial substituents instead of a ring bond. The elongated bonds of axial hyperconjugation (vertical cation) readily undergo 1,2-shifts. When the carbocationic intermediate is horizontally oriented and hyperconjugated to a C-C ring bond, , the 1,2-shift will result in ring enlargement.[61] Analysis of the vertical nepehinyl cation shows that H19 bond length (C19-H19) is 1.12 A, whereas the other axial C-H bonds in the molecule are generally less or equal to 1.10 A. Similarly, in the horizontal cation of the 17βH epimer, the C19-C21 bond is elongated to 1.63 A vs. ~1.55 A of other C-C bonds. The 17αH and 17βH epimers shown in Scheme 4.4 both contain an equilibrium mixture of vertical and horizontal cations, but the vertical cations predominate in the 17αH epimer and are extremely minor in the usual 17βH epimers.

QM calculations suggest that activation energy to convert the nepehinyl cation from to vertical is only 5 kcal/mol versus the 12 kcal/mol when the lupanyl cation undergoes the same transition. This means that lupanyl cation has to overcome a higher energy barrier to go from vertical (rearrangement) to horizontal (ring expansion), thus explaining why Castillo, D. A. and Shan, H. found «1% rearranged products when lupanyl cation was involved. On the other hand, the nepehinyl cation has better access to the unfavored orientation, which accounts for the products of ring expansion found in acid isomerization of nepehine-3β,20-diol.

Nematocyphol synthase from *L. perennis* is the first OSC ever reported to synthesize nematocyphol **1** or any other rearranged lupenols in significant amounts (>5%). NEM is fairly good at controlling rearrangement with deprotonation to make a Δ7 bond. Nematocyphol has also been reported as a minor product of *Ath*PEN6,[59] parallel to my present observation of **1** in AlyPEN6. In both cases, only the Δ 7 rearranged lupenol **1** is detected.

Thus, nematocyphol is known to be part of the product profile of at least two other organisms from the Brasicaceae family, *A. thaliana* and *A. lyrata*. In both cases, nematocyphol is among the minor products of *Aly*PEN6 and *Ath*PEN6, both of them being seco-amyrin synthases. In a previous study by Shibuya *et al.*,[3] it was proposed that in order to perform the Grob fragmentation that permits the formation of seco amyrins, the enzyme needs to have an enlarged active site cavity, which permits the proper alignment of the C13 empty orbital with the C8-C14 bond that is broken during this reaction. This alignment will be possible by allowing the enzyme to accommodate and bend the DE rings slightly upwards, thus favoring Grob fragmentation over elimination to β-amyrin formation.

The suggestion of a slightly enlarged active site cavity around the DE-ring supports the hypothesis previously discussed for the formation of rearranged lupane type products. Formation of the nepehinyl intermediate is thought to be favored in NEM, thus preventing the steric problems that are encountered by rearrangement of lupanyl cation. An enlarged active site cavity, like the one described for *A. thaliana* PEN6 will have enough room to maintain the isopropyl group of the carbocationic intermediate in the correct position (isopropyl group up) in order to start the rearrangement. Consistent with the reasoning that a high ceiling over rings DE is needed for both lupanyl cation rearrangements and seco-

amyrin-type Grob fragmentation, NEM makes a trace of seco-β-amyrin, just as PEN6 makes a small amount of nematocyphol. The shape of the high ceiling evidently controls whether seco-amyrins or rearranged lupenol products predominate.

4.5 Final remarks

Throughout this chapter it is demonstrated that EST mining approaches are useful tools for the discovery of novel OSCs when the genome of an organism is not available; however, this strategy does not provide a complete picture of the variety of triterpenes or number of OSCs in the *Lactuca* family.

Heterologous gene expression used to study *L. perennis* nematocyphol synthase is still a reliable methodology for the analysis of putative OSCs. This work shows the importance of a good approach for OSCs expression, product isolation and characterization in order to get a reliable triterpene product profile. With 38 products that range from 55% to <1%, it is important to have a reliable methods of quantification which in this case were NMR and GC-MS techniques.

Finally, elucidating such a complex product profile provided insights into triterpene biosynthesis in other plant families. OSCs are phylogenetically quite different between *Arabidopsis* and *Lactuca*, but still both organisms have managed to share most of the known triterpene alcohols found in other plants.

The next chapter will discuss strategies for identifying OSCs in *Lactuca* when genomic sequences as well as ESTs are available. Several characterized OSCs from *L. sativa* and *L. serriola* will be analyzed by genome mining strategies

REFERENCES

[1] Truco, M. J.; Reyes-Chin-Wo, S; Lavelle, D.; Kozik, A.; Froenicke, L.; Xu, H.; Michelmore, R. *Lettuce Genome Assembly, Structure and Annotation*, Plant & Animal Genome XXII Conference, San Diego, CA, January 11-15, 2014.

[2] Fazio, C. F. *Ph.D. Dissertation*, Rice University **2006**.

[3] (a) Corey, E. J.; Matsuda, S. P. T.; Bartel, B. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 11628-11632. (b) Kolesnikova, M. D.; Xiong, Q.; Lodeiro, S.; Hua, L.; Matsuda, S. P. T. *Arch. Biochem. Biophys.* **2006**, *447*, 87-95. (c) Herrera, J. B. R.; Bartel, B.; Wilson, W. K.; Matsuda, S. P. T. *Phytochemistry* **1998**, *49*, 1905-1911. (d) Kushiro, T.; Shibuya, M.; Masuda, K.; Ebizuka, Y. *Tetrahedron Lett.* **2000**, *41*, 7705-7710. (e) Kolesnikova, M. D.; Wilson, W. K.; Lynch, D. A.; Obermeyer, A. C.; Matsuda, S. P. T. *Org. Lett.* **2007**, *9*, 5223-5226. (f) Shibuya, M.; Katsube, Y.; Otsuka, M.; Zhang, H.; Tansakul, P.; Xiang, T.; Ebizuka, Y. *Plant Physiol Biochem.* **2009**, *47* (1), 26-30. (g) Ebizuka, Y.; Katsube, Y.; Tsutsumi, T.; Kushiro, T.; Shibuya, M. *Pure Appl. Chem.* **2003**, *75*, 369-374. (h) Xiang, T.; Shibuya, M.; Katsube, Y.; Tsutsumi, T.; Otsuka, M.; Zhang, H.; Masuda, K.; Ebizuka, Y. *Org. Lett.* **2006**, *8*, 2835-2838. (i) Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Kolesnikova, M. D.; Onak, C. S.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2007**, *129*, 11213-11222. (j) Morlacchi, P.; Wilson, W. K.; Xiong, Q.; Bhaduri, A.; Sttivend, D.; Kolesnikova, M. D.; Matsuda, S. P. T. *Org. Lett.* **2009**, *11*, 2627-2630. (k) Fazio, G. C.; Xu, R.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2004**, *126*, 5678-5679. (l) Xiong, Q.; Wilson, W. K.; Matsuda, S. P. T. *Angew. Chem. Intl. Ed.* **2006**, *45*, 1285-1288. (m) Shibuya, M.; Xiang, T.; Katsube, Y.; Otsuka, M.; Zhang, H.; Ebizuka, Y. *J. Am. Chem. Soc.* **2007**, *129*, 1450-1455.

[4] Shibuya, M.; Zhang, H.; Endo, A.; Shishikura, K.; Kushiro, T.; Ebizuka, Y. *Eur. J. Biochem.* **1999**, *266* (1), 302-307.

[5] (a) Michelmore, R.; Kozik, A.; Truco, M. J.; Matviencho, M.; Ochoa1, O.; Van Damme, M.; Lavelle, D.; Lin, H.; Pande, B.; McHale, L.; Sudarshana, P.; Argyris, J.; Ellison, P.; Bradford, K.; Jackson, L.; Kesseli, R. *Eucapiia Leafy Vegetables* **2003**, (eds. Th. J. L. van Hintum, A. Lebeda, D. Pink, J. W. Schut) 131-137. (b) The Arizona Genomics Institute is leading an ongoing effort to sequence expressed sequence tags (ESTs) from both *Lactuca serriola* and *Lactuca sativa* to elucidate the differences between the genomes.

[6] (a) Kuang, H.; Ochoa, O. E.; Nevo, E.; Michelmore, R. W. *Plant J.* **2006**, *47*, 38-48. (b) Kuang, H.; Woo, S. S.; Meyers, B. C.; Nevo, E.; Michelmore, R. W. *Plant Cell.* **2004**, *16*, 2870-2894. (c) Meyers, B. C.; Shen, K. A.; Rohani, P.; Gaut, B. S.; Michelmore, R. W. *Plant Cell* **1998**, *10*, 1833-1846.

[7] Dixon, R. A. *Nature* **2001**, *411*, 843-847.

[8] Doireau, P.; Quemerais, D.; Verger, A. *Biochem. Soc. Trans.* **1983**, *11*, 595-595.

[9] Shan, H.; Wilson, W. K.; Castillo, D. A.; Matsuda, S. P. T. *Org. Lett.* **2015**, *17*, 3986-3989.

[10] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision B.01, Gaussian, Inc., Wallingford CT, 2009.

[11] Guo, L.-W.; Wilson, W. K.; Shackeleton, C. H. L. Unpublished results.

[12] Kolesnikova, M. D.; Obermeyer, A. C.; Wilson, W. K.; Lynch, D. A.; Matsuda, S. P. T. *Org. Lett.* **2007**, *9*, 2183-2186.

[13] Morlacchi, P.; Wilson, W. K.; Xiong, Q.; Bhaduri, A.; Sttivend, D.; Kolesnikova, M. D.; Matsuda, S. P. T. *Org. Lett.* **2009**, *11*, 2627-2630.

[14] Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Kolesnikova, M. D.; Onak, C. S.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2007**, *129*, 11213-11222.

[15] Shan, H.; Segura, M. J. R.; Wilson. W. K.; Lodeiro, S.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2005**, *127*, 18008-18009.

[16] Lodeiro, S.; Wilson. W. K.; Shan, H.; Matsuda, S. P. T. *Org. Lett.* **2006**, *8*, 439-442.

[17] Xiong, Q.; Wilson, W. K.; Matsuda, S. P. T. *Angew. Chem Int. Ed. Engl.* **2006**, *45*, 1285-1288.

[18] Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Ivanova, Y.; Smith, M. L.; May, G. S.; Matsuda, S. P. T. *Org. Lett.* **2009**, *11*, 1241-1244.

[19] Castillo, D. A.; Kolesnikova, M. D.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2013**, *135*, 5885-5894.

[20] Cao, D.; Su Y. L.; Yang J. S. *Acta Pharmacol. Sin.* **1991**, *27*, 445-451.

[21] (a) Laird, W.; Spring, F. S.; Stevenson, R. J. Chem. Soc. **1961**, 2638−2642. (b) Agata, I.; Corey, E. J.; Hortmann, A. G.; Klein, J.; Proskow, S.; Ursprung, J. J. *J. Org. Chem.* **1965**, *30*, 1698−1710. (c) Chivers, H.; Corbett, R. E.; Mitchell, R. E. M. *J. Chem. Soc. C* **1966**, 1814−1816.

[22] (a) Xu, Z.; Wang, Q.; Xhao, J. *Zhomgcaoyao* **1983**, *14*, 49-51. (b) Kawanishi, K.; Hashimoto, Y.; Qiang, W.; Zhenwen, X. *Phytochemistry* **1985**, *24*, 2051-2054.

[23] Goad, L.J.; Akihisa, T.; Analysis of Sterols; Blackie (Chapman & Hall): London, **1997**.

[24] Ageta, H.; Shiojima, K.; Masuda, K.; Lin, T. *Tetrahedron Lett.* **1981**, *22*, 2289-2290.

[25] (a) Ref. 24. (b) Kaith, B. S. *Int. J. Pharmacogn.* **1996**, *34*, 73-75. (c) Rahman, A. *Indian J. Chem., Sect B* **2004**, *43B*, 2446-2451.

[26] Ames, T. R.; Beton, J. L.; Bowers, A.; Halsall, T. G.; Jones, E. R. H. *J. Chem. Soc.* **1954**, 1905-1919.

[27] Shibuya, M.; Xiang, T.; Katsube, Y.; Otsuka, M.; Zhang, H.; Ebizuka, Y. *J. Am. Chem. Soc.* **2007**, *129*, 1450-1455.

[28] Talapatra, S. K.; Sengupta, S.; Talapatra, B. *Tetrahedron Lett.* **1968**, 5963-5968.

[29] Ames, T. R.; Halsall, T. G.; Jones, E. R. H. *J. Chem. Soc.* **1951**, 450-457.

[30] Chatterjee, A.; Kundu, A. B.; Chakrabortty, T.; Chandrasekharan, S. *J. Chem. Soc. Commun.* **1968**, 418-419.

[31] Itoh, T.; Tamura, T.; Matsumoto, T. *Lipids* **1976**, *11*, 434-441.

[32] Su, Y.; Tachibana, S.; Sumimoto, M. *Mokuzai Gakkaishi* **1986**, *32*, 190−202.

[33] Gonzales, A. G.; Gutierrez, Jerez, F.; Luque Escalona, M. *An. Quim* **1973**, *69*, 921−928.

[34] (a) Herrera, J. B. R.; Bartel, B.; Wilson, W. K.; Matsuda, S. P. T. *Phytochemistry* **1998**, *49*, 1905-1911. (b) Segura, M. J. R.; Meyer, M. M.; Matsuda, S. P. T. *Org. Lett.* **2000**, *2*, 2257-2259.

[35] Yokoyama, Y.; Niriyama, Y.; Tsuyuki, T.; Takahashi, T. *Bull. Chem. Soc. Jpn.* **1981**, *54*, 234-239.

[36] Mills, J. S. *J. Chem. Soc.* **1956**, 2196-2202.

[37] (a) Lahey, F. N.; Leeding, M. V. *Proc. Chem. Soc.* **1958**, 2196-2202. (b) Fukuoka, M.; Natori, S. *Chem. Pharm. Bull.* **1972**, *20*, 974-979.

[38] Bauer, S.; Schulte, E.; Their, H. P. *Eur. Food Res. Technol.* **2004**, *219*, 487–491.

[39] Goad, L.J.; Akihisa, T.; Analysis of Sterols; Blackie (Chapman & Hall): London, **1997**.

[40] (a) Ageta, H.; Shiojima, K.; Masuda, K.; Lin, T. *Tetrahedron Lett.* **1981**, *22*, 2289-2290. (b) Kawanishi, K.; Hashimoto, Y.; Qiang, W.; Zhenwen, X. *Phytochemistry* **1985**, *24*, 2051-2054.

[41] Bischof, B.; Jeger, O.; Ruzicka, L. *Helv. Chim. Acta* **1949**, *32*, 1911-1921.

[42] Brooks, C. J. W. *Chem. Ind.* **1953**, 1178.

[43] Akihisa, T.; Yasukawa, K.; Kimura, Y.; Takase, S.-I.; Yamanouchi, S.; Tamura, T. *Chem. Pham. Bull.* **1997**, *45*, 2016-2023.

[44] (a) Kitajima, J.; Arai, M. Tanaka, Y. *Chem. Pham. Bull.* **1994**, *42*, 608-610. (b) Kiyotani, T.; Kitajima, J.; Tanaka, Y.; Ageta, H. *Acta Cryst.* **1996**, C*52*, 2024-2026.

[45] Bartlett, W.R.; Johnson, W.S.; Plummer, M.S.; Small, V.R. *J. Org. Chem.* **1990**, *55*, 2215-2224.

[46] Kushiro, T.; Shibuya, M.; Masuda, K.; Ebizuka, Y. *Tetrahedron Lett.* **2000**, *41*, 7705-7710.

[47] Susunaga, G. S.; Siani, A. C.; Pizzolatti, M. G.; Yunes, R. A.; Delle Monache, F. *Fitoterapia* **2001**, *72*, 709-711.

[48] Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Kolesnikova, M. D.; Onak, C. S.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2007**, *129*, 11213-11222.

[49] Kirby, J.; Romanini, D. W.; Paradise, E. M.; Keasling, J. D. *FEBS J.* **2008**, *275*, 1852-1859.

[50] Kushiro, T.; Shibuya, M.; Ebizuka,Y. *Eur. J. Biochem.* **1998**, *256*, 238-244.

[51] Kushiro, T.; Shibuya, M.; Ebizuka, Y. Molecular Cloning of Oxidosqualene Cyclase cDNA from Panax ginseng: The Isogene that Encodes β-Amyrin Synthase. International Symposium on Natural Medicines; **1998**, Kyoto Japan.

[52] Huang, L.; Li, J.; Ye, H.; Li, C.; Wang, H.; Liu, B.; Zhang, Y. *Planta* **2012**, *236*, 1571-1581.

[53] Saimaru, H.; Orihara, Y.; Tansakul, P.; Kang, Y.H.; Shibuya, M.; Ebizuka, Y. *Chem. Pharm. Bull.* **2007**, *55*, 784-788.

[54] (a) Tansakul, P.; Shibuya, M.; Kushiro, T.; Ebizuka, Y. *FEBS Lett.* **2006**, *580*, 5143-5149. (b) Hu, W.; Liu, N.; Tian, Y.; Zhang, L. *Biomed. Res. Int.* **2013**, 285740.

[55] Niu, Y.; Luo, H.; Sun, C.; Yang, T. J.; Dong, L.; Huang, L.; Chen, S. *Gene* **2014**, *533*, 295-303.

[56] Luo, H.; Sun, C.; Sun, Y.; Wu, Q.; Li, Y.; Song, J.; Niu, Y.; Cheng, X.; Xu, H.; Li, C.; Liu, J.; Steinmetz, A.; Chen, S. *BMC Genomics* **2011**, *12* (Suppl. S5).

[57] Sawai, S.; Uchiyama, H.; Mizuno, S.; Aoki, T.; Akashi, T.; Ayabe, S.; Takahashi, T. *FEBS Lett.* **2011**, *585*, 1031-1036.

[58] Castillo, D. A.; Kolesnikova, M. D.; Matsuda, S. P. T. *J. Am. Chem. Soc*. **2013**, *135*, 5885-5894.

[59] Castillo, D. A.; *Ph.D. Dissertation*, Rice University **2014**.

[60] Dorianne A. Castillo, Matias I. Kinzurik and Hui Shan, unpublished data.

[61] Matsuda, S. P. T.; Wilson, W. K.; Xiong, Q. *Org. Biomol. Chem.* **2006**, *4*, 530-543.

CHAPTER 5

Genome and EST Mining Approaches to Elucidate Triterpene Biosynthesis in Lettuce

This chapter describes my work towards the analysis and characterization of novel oxidosqualene cyclases in lettuce. The many lettuce species from the *Lactuca* genus, which is a member of the Asteraceae (or Compositae) family, are asterids. The increasing availability of the *Lactuca sativa* genome during this project helped to give a more complete picture of triterpene biosynthesis in lettuce. *Lactuca sativa* is the first plant from this family to have its genome available and can be used as a model organism for studying triterpene biosynthesis in Compositae.

In this chapter, I present the product profiles of five oxidosqualene cyclases from *Lactuca sativa* and *Lactuca serriola* (Table 5.1). Although these OSCs were previously reported by Gia Fazio,[1] a vast improvement in our laboratory's analytical methodologies helped me to expand our knowledge about the product profiles of these cyclases, hence the possibility to expand our understanding of OSC catalysis in Compositae.

Table 5.1. Summary of the *Lactuca* OSCs expressed for this study.

Plasmid	GeneBank ID	Major Product	Species
pGCF14.3	BU000828	tirucalla-8,24-dien-3β-ol	*L. serriola*
pGCF15.3	BQ869693	α-amyrin	*L. sativa*
pGCF16.9	BQ997802	taraxasterol	*L. serriola*
pGCF17.6	BU003435	lupeol	*L. serriola*
pMMG11.2	BU003176	[nematocyphol]*	*L. serriola*
pGCF19.6	BU008606	cycloartenol	*L. serriola*

* No product was observed from EHY41[pMMG11.2] or RXY6[pMMG11.2]

Product numbering for this chapter, as in chapter 3, is assigned in descending order of amount, starting with **1** for the major product of each enzyme. Thus, the same triterpene structure may have different numbers in different sections of this chapter. Another set of

numbers is used for the plant extracts. Additional information, sequences and spectral data can be found in the Appendix.

5.1 Background and Previous Work

The Compositae or Asteraceae family, where *Lactuca* is a member genus, is one of the largest and most diverse families of flowering plants, comprising almost one-tenth of all known Angiosperm species.[2] The Compositae family contains over 40 economically important species like lettuce, which generates over 1.5 billion dollars in revenue as a group.[3]

In the Cichorieae (or Lactuceae) tribe of Asteraceae, Lactuca has historically been the best genetically characterized genus.[4] Disease is an important factor in lettuce production and has motivated studies of genetic variance at disease resistance loci.[5] Diverse secondary metabolites can contribute resistance to herbivores in plant systems. The defense role of triterpenes, which are abundant in lettuce, is mostly unexplored.

Sequencing of the *Lactuca sativa* genome was announced in 2013,[6a] and the analysis tools were made accessible in the Lettuce Genome Resource where several versions of the data were displayed in GBrowse and could be used to compare with other plant genomes. In addition, lettuce genes were assigned a KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathway, which made my search for putative OSCs easier within the published genomic contigs.[7] Now genomes of other Compositae species, including horseweed, sunflower, and globe artichoke, are at various stages of release.[6b]

Investigation into OSCs in the *Lactuca* species began in the Matsuda Laboratory around 2005 with the analysis of ESTs by Gia Fazio.[1] Large collections of ESTs have long been available for five *Lactuca* species and other Cichorieae.[4b] However, the ESTs-only approach does not establish how many putative OSCs *Lactuca* has. As expected, most *Lactuca* OSCs besides cycloartenol synthase (*L. serriola* BU008606) have no close counterparts in the *Arabidopsis thaliana* genome.

Previous work of heterologous expression of these cyclases in the Matsuda Research group was done mainly by Gia Fazio; followed by experiments on the various cyclases by Carl Onak, Caroline V. McNeil, Aparna Bhaduri, Jazmin Godoy and Matias

Kinzurik. These experiments were afflicted by yeast strain problems and low quality NMR and GC-MS analysis. In the following sections, I will present a more complete product profile of each of the cyclases mentioned above, and I will discuss the implications of genome mining for the discovery of OSCs that seem characteristic of the *Lactuca* family.

5.2 Experimental Procedures

5.2.1 Phylogenetic Analysis of *Lactuca sativa* Genome

Genomic DNA sequences and computer spliced sequences were downloaded from the Lettuce Genome Resource[7,8] using the enzyme commission number (EC number) search provided by the database. The sequence of EC 5.4.99.8 (cycloartenol synthase) from *A. thaliana* was used to look for a match in the genomic sequences and/or computer spliced sequences deposited in the Lettuce Genome Resource. Forty pieces of genomic sequence of various lengths were downloaded and assembled using SeqMan,[9] these sequences were assembled in 19 different contigs and manually spliced using as references ESTs from different *Lactuca* species that were available in the NCBI Bio Systems database,[10] and available sequences from Gia Fazio's *Lactuca* OSCs.[1]

Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 6.[11] Multiple alignments of nucleotide sequences were generated from amino acid sequence with Clustal W Method using default parameters. Maximum likelihood trees were constructed using MEGA version 6.[11]

5.2.2 Plasmids and Yeast Strain Construction

5.2.2.1 PCR-Mutagenesis of pGCF18.5: Construction of pMMG11.2

Plasmid pGCF18.5 contains the putative OSC BU003176 from *Lactuca serriola*, and it was described by Gia Fazio[1] as a pseudogene because her clone had two nucleotides (TC) at position 1834 missing that will create an early truncation upon expression, resulting in a putative OSC without the C-terminus. Since the genome of *Lactuca serriola* is not

available, I looked into the genomic sequence to see if these missing nucleotides were the product of evolution or a PCR mistake during the creation of the clone *BU003176*.

Alignment of BU003176 with its close homologs (Figure 5.1) from *Lactuca sativa* and *Lactuca perennis* suggested that the absence of these nucleotides could be the product of a PCR mistake, since both closest homologs do have the missing nucleotides in the conflict region and, when they are inserted in BU003176, its amino acid sequence has a high percentage similarity to the homologs in *L. sativa* and *L. perennis* (nematocyphol synthase).

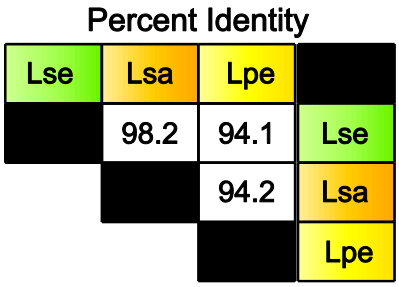


Figure 5.1. Table of amino acid percentage identity values for corrected sequence of pGCF18.5 compared with its homologs. Green corresponds to *L. serriola* BU003176, orange to *L. sativa* homologs and yellow to *L. perennis* nematocyphol synthase.

The original BU003176 clone was obtained from Arizona Genome Institute AGI (www.genome.arizona.edu) and sequenced with pertinent primers to confirm the ~2.3 kbp cDNA sequence. *Sal I* and *Not I* restriction sites were added by PCR amplification, and the amplicon was subcloned into the pGEM-T vector system.

DNA mutagenesis of BU003176 was conducted as described in Chapter 2. It was done using the following primers: (1) a forward primer (5'-CCCGTCGACAAAATGTGGGAGTTAAAG-3') that starts amplification at the ATG start site, (2) a reverse primer (5'-CCCGATATCGTATTCCACTCAAAATTGAAAGTTC-3') that introduced the TC nucleotides missing for this clone and (3) an *EcoRV* restriction site that allowed substitution with the piece with the point mutation into the BU003176 pGEMT-vector predigested plasmid using the restriction site alignment.

The correct construct (with the TC nucleotides added) was then digested with *SalI* and *NotI* restriction enzymes and gel purified, and cloned into the yeast expression vector pRS426GAL carrying the *URA3* markers. The resultant plasmid containing the ~2.3 kbp gene was named pMMG11.2 and it was used to transform the EHY41 and RXY6 yeast strains; however, it did not show any sign of cyclase expression in either strain. This result shows some similarities to the case of *Aly*PEN8, but in the case of pGCF18.5 there is evidence (in the form of ESTs) of the presence of this OSC in the plant. Other possibilities for lack of expression in yeast could be that this particular OSC is susceptible to yeast proteases. At this point, efforts were abandoned to express pGCF18.5 from *L. serriola* (pMMG11.2). Instead, plasmid pMMG12.3 was constructed from an EST (DW077158) of *L. perennis* to

5.2.2.2 Yeast Strain Construction for the Plasmids pGCF14.3, pGCF15.3, pGCF16.9, pGCF17.6 and pGCF19.6

The plasmids listed below in Table 5.2 were used to transform *Saccharomyces cerevisiae* strain RXY6. RXY6 transformants were selected on synthetic complete medium lacking uracil, solidified with 1.5% agar, and supplemented with 2% glucose, 13 mg/L hemin chloride, 20 mg/L ergosterol, and 5 g/L Tween 80.

In vitro expression of pGCF14.3 (tirucalla-8,24-dien-3β-ol synthase) proved to be challenging; it was repeated numerous times, at various pH and scales, but it never produced meaningful amounts of products. The effort of providing a product profile using *in vitro* expression was abandoned and instead I present my joint work with Matias Kinzurik in an effort to elucidate the product profile of this particular enzyme.

Yeast transformation of pGCF14.3 was done by Matias Kinzurik in the strain EHY41 using the lithium acetate method. Further purification and analysis was done either under my guidance or directly by me.

5.2.3 *In vitro* and *in vivo* Experiments of Putative *Lactuca OSCs*

Cyclases shown in Table 5.2 (except A and E), were previously reported by Gia Fazio. [1] Nevertheless, improved analytical techniques allowed me to present in this work an improved picture and understanding of the product profile for each one of these cyclases, which is important for the overall analysis of the metabolic triterpene capability of the *Lactuca* family. GC-MS data presented in Figures 5.2 to 5.4 follow the compound numbering presented in the results section of this chapter (Table 5.7 to 5.12).

Table 5.2. Plasmids and Yeast Strains corresponding to the six expressed *Lactuca* OSCs expressed.

	Plasmid	Expression vector	Yeast Strain	Reported major products	Organism
A	pGCF14.3	pRS426GAL	EHY41	tirucalla-8,24-dien-3β-ol	*L. serriola*
B	pGCF15.3	pRS426GAL	RXY6	mixed amyrin	*L. sativa*
C	pGCF16.9	pRS426GAL	RXY6	mixed taraxasterol	*L. serriola*
D	pGCF17.6	pRS426GAL	RXY6	lupeol	*L. serriola*
E	pMMG11.2	pRS426GAL	RXY6	nematocyphol	*L. perennis*
F	pGCF19.6	pRS426GAL	RXY6	cycloartenol	*L. serriola*

Table 5.3. *In vitro* and *in vivo* experiments of *Lactuca* OSCs.

	Yeast Strain Constructs	Culture size	pH*	Cell pellet g	Crude mg
A	EHY41[pGCF14.3]	4 L	NA	80	200
B	RXY6[pGCF15.3]	2 L	6.2	26.3	208
C	RXY6[pGCF16.9]	2 L	7	23.74	163
D	RXY6[pGCF17.6]	4 L	6.4	53	177
E	RXY6[pMMG11.2]	2 L	7	---	---
F	RXY6[pGCF19.6]	2 L	7	26	240

* NA non applicable

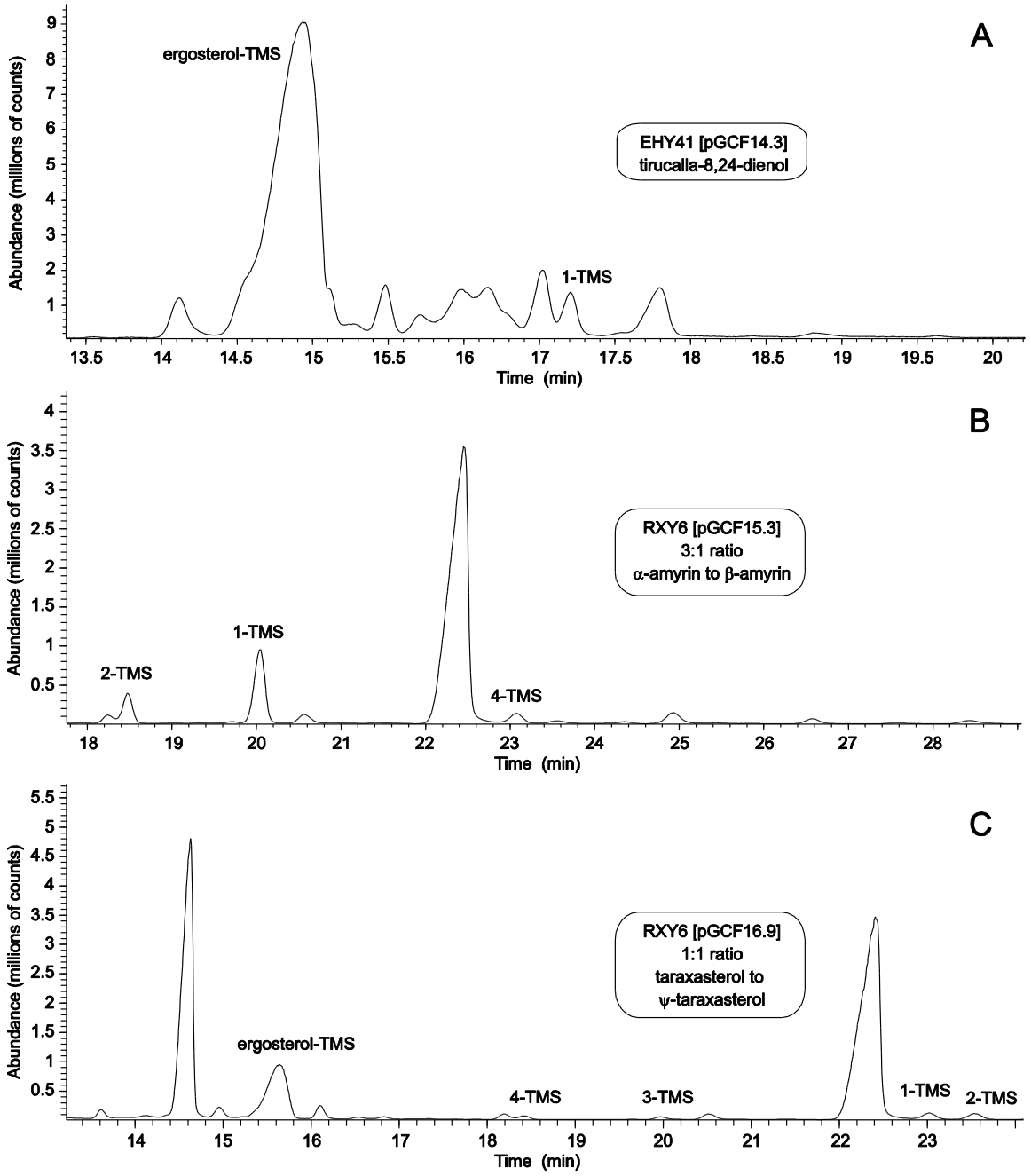


Figure 5.2. A: GC-MS chromatograms of the crude extract of the 4 L EHY41[pGCF14.3] experiments. B: GC-MS chromatograms of the crude extract of the 2 L RXY6[pGCF15.3] experiments. C: GC-MS chromatograms of the crude extract of the 2 L RXY6[pGCF16.9] experiments.

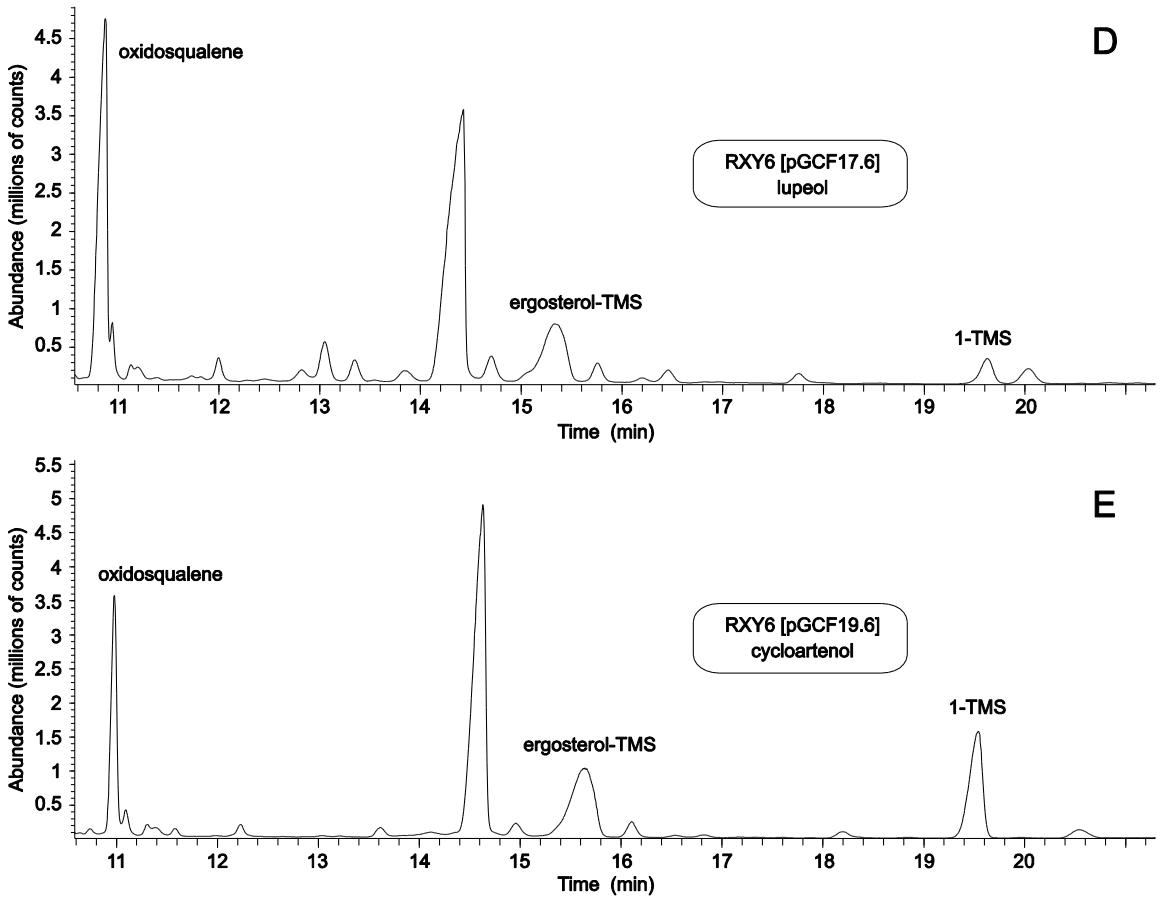


Figure 5.3. D: GC-MS chromatograms of the crude extract of the 2 L RXY6[pGCF17.6] experiments. E**:** GC-MS chromatograms of the crude extract of the 2 L RXY6[pGCF19.6] experiments.

The samples were purified by SPE (solid phase extraction). Each crude was dissolved in 2 mL of hexanes, and loaded onto a 5g SiO_2 gel SPE column (Phenomenex). 60 x 8 mL fractions were collected using diethyl ether (Et2O)/ hexanes gradient shown in Table 5.4.

Fractions containing the triterpene alcohols and triterpene diols (usually but not limited to fractions 19 to 33 for triterpene alcohols and fractions 45 to 50 for triterpene diols) were combined to yield a combine sample of triterpene alcohols and one of triterpene diols (Table 5.5). These samples were used for GC-MS and NMR analysis. Proton NMR and HSQC spectra, were collected for these samples.

Combined fractions containing triterpene alcohols were subjected to small scale saponification to eliminate triglycerides and other impurities in order to obtain cleaner NMR spectra.

Table 5.4. SPE elution gradient of *Lactuca in vitro* experiments.

Fraction	% Et_2O in Hexanes	Fraction	% Et_2O in Hexanes
1-2	0	27-32	10%
3-6	1%	33-38	20%
7-12	2%	39-46	30%
13-18	3%	47-54	50%
19-26	5%	55-60	100%

Table 5.5. Condensed analysis of *Lactuca* OSCs.

	Yeast Strain Constructs	SPE Fx triterpene alcohols	Weight mg	SPE Fx triterpene diols	Weight mg	Saponification of triterpene alcohols Fx
A	EHY41[pGCF14.3]	22-28	10	---	---	---
B	RXY6[pGCF15.3]	21-33	7	43-51	~ 2	●
C	RXY6[pGCF16.9]	22-32	5.5	44-47 50-51	~ 2	●
D	RXY6[pGCF17.6]	21-32	3.5	42-45	~ 2	●
E	RXY6[pMMG11.2]	---	---	---	---	---
F	RXY6[pGCF19.6]	20-33	18.8	---	---	●

5.2.4 Extraction and Characterization of Lettuce Plant Tissue

Lettuce tissue was harvested from two *Lactuca sativa* plants donated by a goat farm in the Houston area. In addition to this material, seeds from *Lactuca sativa* were purchased from Amazon.com in order to complete the analysis of all parts of this organism.

The plant tissue was divided into the following categories: flowers, roots, leaves, stem and seeds. Each part was weighted and extracted according to Shan, H. *et al.*[12]

Table 5.6 illustrates the final weight of fresh tissue for each part of the plant, NLSs after extraction, the weight of the final crude and the amount that was taken for further analysis, including GC-MS and NMR.

Plant tissue was crushed using a mortar and pestle and soaked in 10 mL of 1:1 methylene chloride/ methanol per gram of tissue for a day with gentle stirring. The methylene chloride in the samples was removed by rotatory evaporation, leaving behind a crude extract.

After centrifugation, the methanolic extracts were saponified with 10% KOH (w/v) in 80% EtOH (v/v). Following saponification, the supernatant was decanted, and the methanol was removed by rotatory evaporation. The aqueous layer was then extracted with 8 x 50 mL hexanes. The combined hexanes layers were concentrated to dryness by rotatory evaporation in a round-bottomed flask, then transferred to pre-weighted scintillation vials and evaporated under nitrogen. The crude non-saponifiable lipids (NSLs) were weighed and analyzed by GC-MS and NMR spectroscopy. Further purification was done following the procedures outlined in Chapter 2.

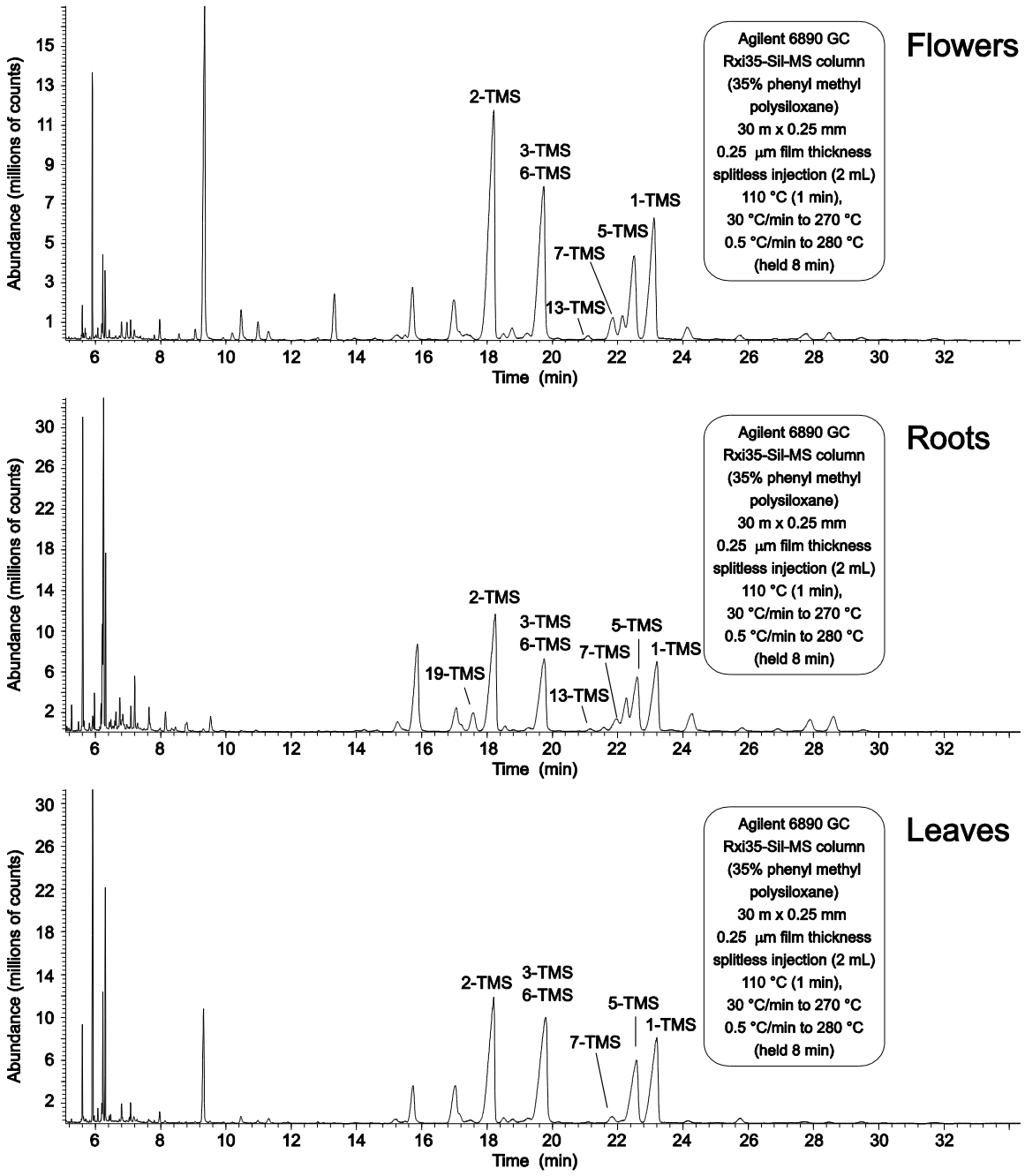


Figure 5.4. GC-MS chromatograms presented in this figure are those of the NSLs of Flowers, Roots and Leaves crude extracts from Lettuce (*Lactuca sativa*).

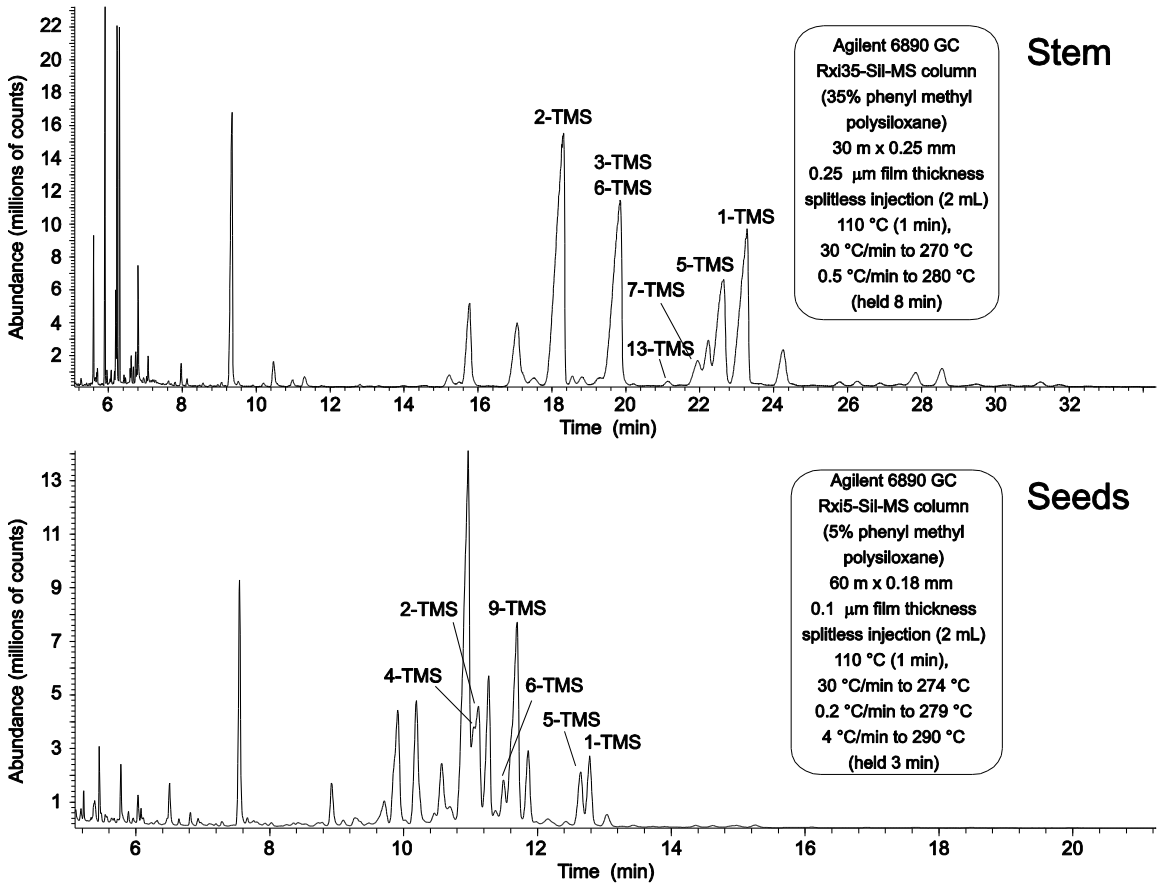


Figure 5.5. GC-MS chromatograms presented in this figure are those of the NSLs of Stem and Seeds crude extracts after saponification from Lettuce (*Lactuca sativa*).

Table 5.6. Condensed information from plant tissue experiments.

Plant part	Weight (g)	Crude of NSLs (mg)	NMR sample (mg)
Flower	26	608	7
Roots	65	148	12
Leaf	66	1036	11
Stem	120	1032	8
Seeds	10	500	30

5.2.5 Extraction of Curcurbitadienol as a Standard from Watermelon Seeds [24]

Eight grams of watermelon seeds (*Citrullus lanatus var. lanatus*) were crushed with mortar and pestle and soaked in 100 mL of hexanes for two days with gentle stirring. The hexanes in the sample was removed by rotatory evaporation, leaving behind ~5 g of crude extract.

8 mL of water, 32 mL of MeOH and 4 g of KOH was added to the sample (10% KOH (w/v) in 80% EtOH) and incubated at 70 ˚C for 3 h. Following saponification, the excess ethanol was removed by rotatory evaporation. The aqueous layer was then extracted 6 x 50 mL with hexanes. The combined hexanes layers were concentrated to dryness by rotatory evaporation in a round-bottomed flask before being transferred to a pre-weighed scintillation vial and evaporated under nitrogen. The crude NSLs weighed 184 mg, and were analyzed by GC-MS (Figure 5.6).

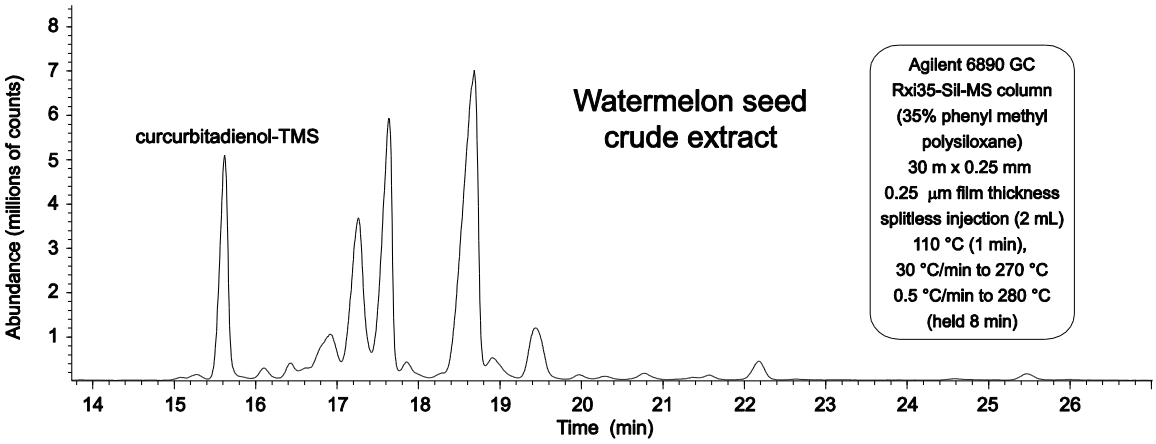


Figure 5.6. GC-MS chromatogram presented in this figure is of the NSLs of Watermelon seed crude extracts after saponification.

The crude NSLs were subsequently dissolved in 2 mL of hexanes, and loaded onto a 5g SiO_2 gel SPE column (Phenomenex). 60 x 8 mL fractions were collected using diethyl ether (Et_2O)/ hexanes gradient (0%, 1%, 2%, 3%, 5%, 10%, 20%, 30%, 50%, and 100% diethyl ether).

Fractions containing the triterpene alcohols (Fx 25 to 31) and triterpene diols (Fx 45 to 51) were combined to yield 4 mg of triterpene products, this sample was used for GC-MS and NMR analysis in order to establish the product profile of triterpene alcohols minor products in watermelon seeds.

Fractions 21 to 24 were combined to get good GC-MS and NMR data for curcurbitadienol since that product was the major component of those fractions.

5.3 Results

5.3.1 Manual Splicing of Genomic Sequence from *Lactuca sativa*: Phylogenetic Relationship of *Lactuca sativa* OSCs

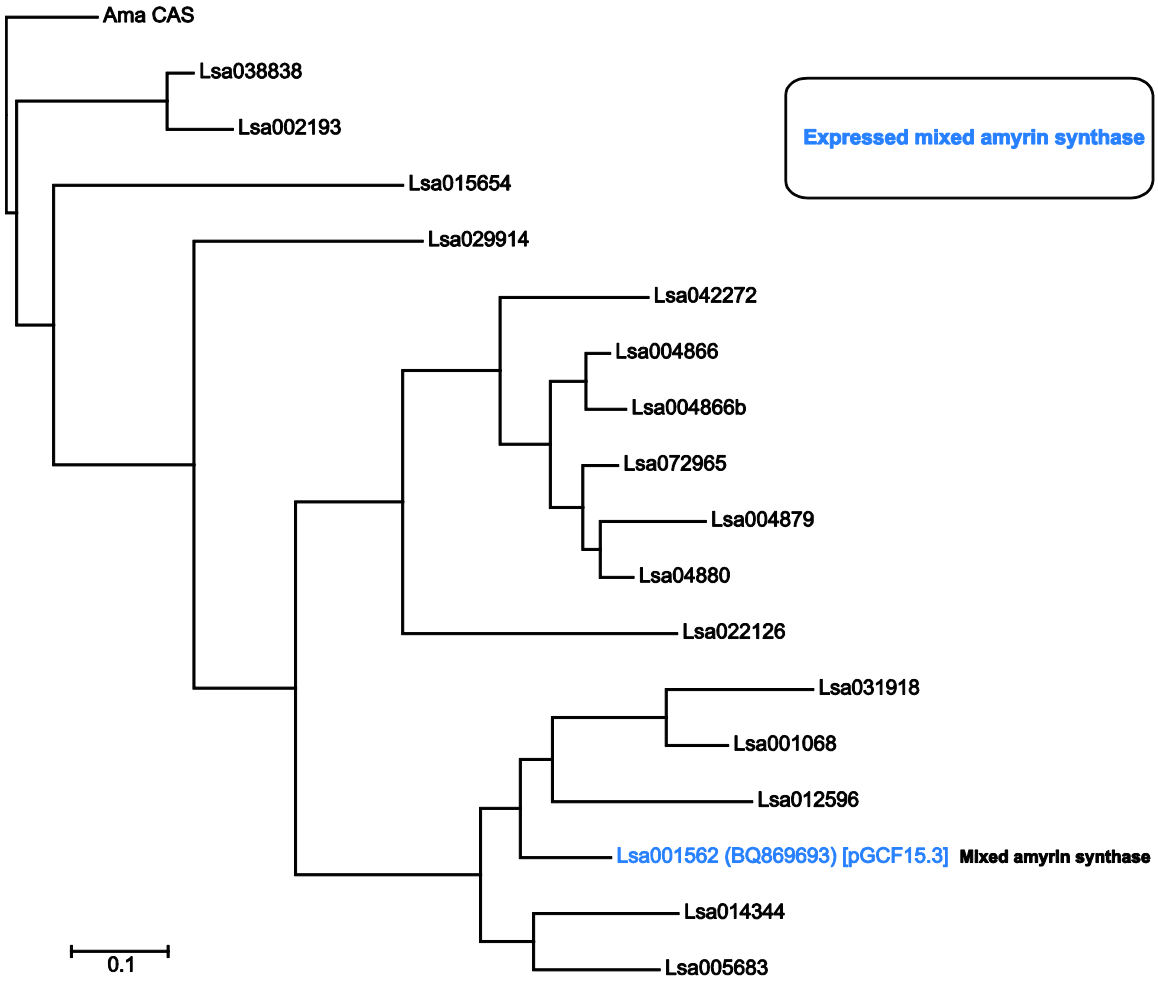


Figure 5.7. Phylogenetic tree of *L. sativa* OSCs derived from the initial manual splicing of the genome. The tree was constructed by maximum likelihood using MEGA version 6.[11] Multiple sequence alignments were generated in MegAlign from amino acid sequences with Clustal W Method using default parameters. After experimental analysis, the gene in blue turned out to be a mixed amyrin synthase.

Extensive analysis of genomic sequence from the *Lactuca sativa* genome generated 17 putative OSCs. As explained in the experimental section of Chapter 5, each cyclase was manually spliced using as a guide either available ESTs from Lettuce (http://blast.ncbi.nlm.nih.gov/Blast.cgi), or *Lactuca serriola* OSCs sequences that were assembled by Fazio.[1]

After manual splicing, each putative OSC was analyzed by comparing a protein query versus translated sequence (tBLASTn). This search was performed against nucleotide collections, "Asterids" database, each query was covered 99%.

Figures 5.7 and 5.8 show the phylogenetic results and the active site sequence alignment for the various putative OSCs of *L.sativa*.

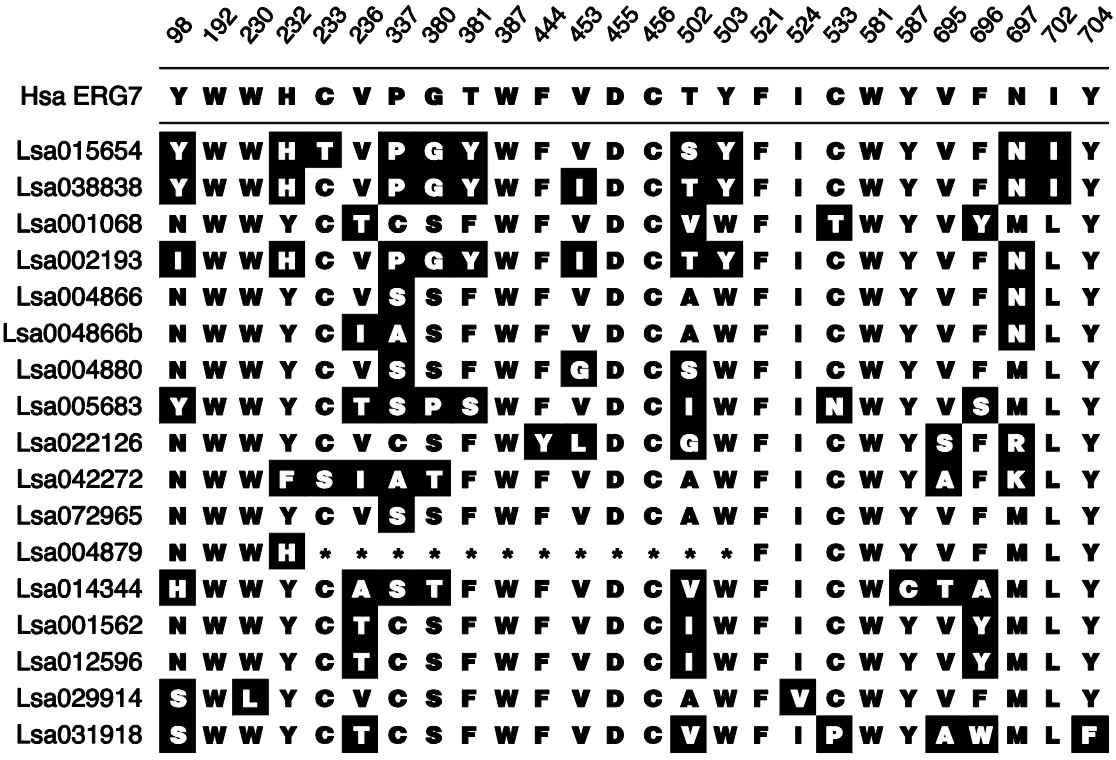


Figure 5.8. Comparison of active-site residues in human LAS[42] and corresponding residues of *Lactuca sativa* OSCs.

5.3.2 *In vivo* and *in vitro* Product Profile of *Lactuca* OSCs from *Lactuca sativa* and *Lactuca serriola*

5.3.2.1 *In vivo* Product Profile of Tirucalla-8,24-dien-3β-ol Synthase

Tirucalla-8,24-dien-3β-ol synthase (BU000828) from *L. serriola* was named after its major product (**1**). This gene was cloned by Gia Fazio; Caroline V. McNeil and Jazmin Godoy performed heterologous expression of this cyclase, so they provided a preliminary product profile from which the foundation presented in this document comes from.

In this work, I present preliminary results of the *in vivo* product profile of this enzyme. Tirucalla-8,24-dien-3β-ol account for 70% of the total of products that this enzyme can make. In addition to tirucalla-8,24-dien-3β-ol (**1**), this cyclase makes at least 3 other triterpene alcohols, and 8% of other minor products.

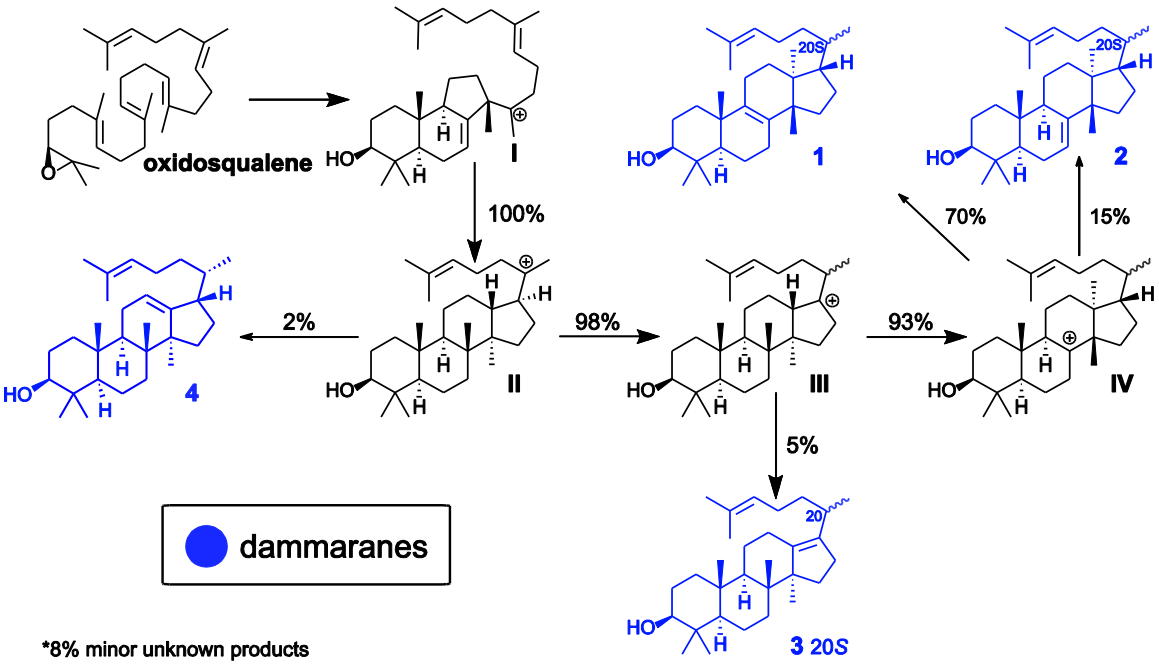
The enzyme mechanism and product profile are presented in Scheme 5.1 and a summary of its products is presented in Table 5.7.

Product profiles of triterpene alcohols were determined as described by Castillo *et al.*[56] Ratios for the triterpene-monoalcohols were determined using HSQC spectrum of EHY41[pMMG14.3] Fractions 22-28.

Table 5.7. Summary of analytical evidence for tirucalla-8,24-dien-3β-ol synthase products identified in the 4 L *in vivo* experiment.

Product	Product #	GC-MS*	NMR Evidence	Reference
tirucalla-8,24-dien-3β-ol	1	A	HSQC, 1D	13
tirucalla-7,24-dien-3β-ol	2	A	HSQC, 1D	14
isotirucallol	3	A	HSQC, 1D	15
20*S*-dammara-12,24-dienol	4	A	HSQC, 1D	16

* GC-MS evidence using Method A, Chapter 2.



Scheme 5.1. Mechanism of the formation of *L. serriola* tirucalla-8,24-dien-3β-ol synthase products.

5.3.2.2 *In vitro* Product Profile of Mixed Amyrin Synthase

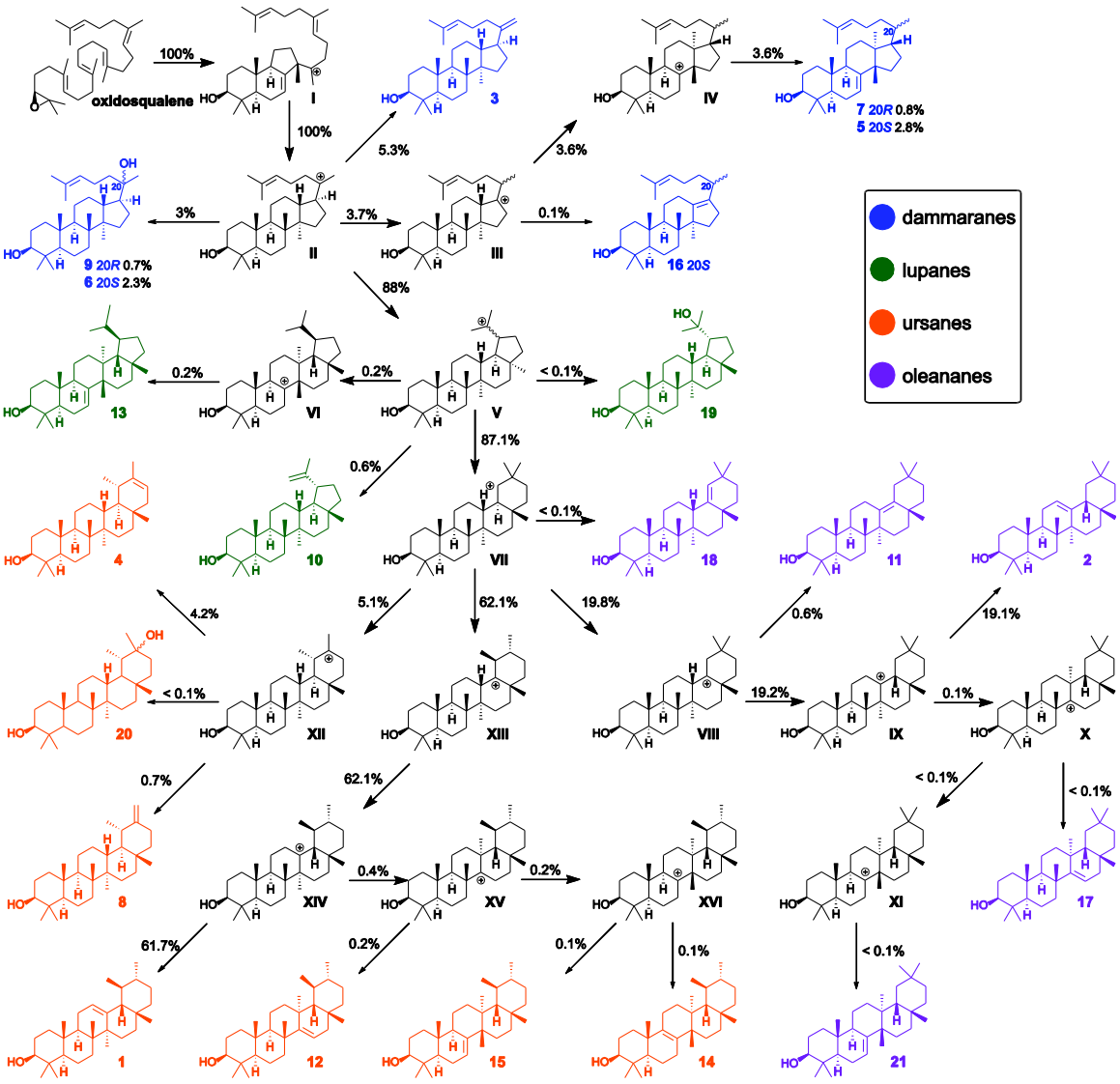
L. sativa mixed amyrin synthase (BQ869693) was cloned by Gia Fazio.[1] This cyclase was inadvertently reported as giving a 3:1 ratio of β-amyrin to α-amyrin, along with 7 minor products. In this work, I report that this enzyme makes a 3:1 ratio of α-amyrin to β-amyrin as well as 19 minor products. The product profile of this enzyme has a variety of tetra and pentacycles that include dammarenes, lupanes, ursanes and oleananes; however, it clearly favors the mechanistic pathway to ursanes.

The enzyme mechanism and product profile are presented in Table 5.8 and Scheme 5.2. As with all cyclases in this work, product profiles were determined as described in Chapter 3. Ratios for the triterpene-monoalcohols and diols were determined using an HSQC spectrum of RXY6[pGCF15.3] Fractions 21-33.

Table 5.8. Summary of analytical evidence for *L. sativa* mixed amyrin synthase products identified in the 2 L *in vitro* experiment.

Product	Product #	GC-MS*	NMR Evidence	Reference
α-amyrin	1	A	HSQC, 1D	17
β-amyrin	2	A	HSQC, 1D	18
dammara-20,24-dienol	3	A	HSQC, 1D	19
ψ-taraxasterol	4	A	HSQC, 1D	20
tirucalla-7,24-dien-3β-ol	5	A	HSQC, 1D	14
(20*S*)-dammarenediol	6	A	HSQC, 1D	21
butyrospermol	7	A	HSQC, 1D	19
taraxasterol	8	A	HSQC, 1D	20
(20*R*)-dammarenediol	9	A	HSQC, 1D	21
lupeol	10	A	HSQC, 1D	22
δ-amyrin	11	A	HSQC, 1D	23
isoursenol	12	A	HSQC, 1D	24
nematocyphol	13	A	HSQC, 1D	25
isobauerenol	14	A	HSQC, 1D	26
bauerenol	15	A	HSQC, 1D	27
isotirucallol	16	A	HSQC, 1D	15
taraxerol	17	A	HSQC, 1D	28
germanicol	18	A	HSQC, 1D	29
lupane-3β,20-diol	19	A	HSQC, 1D	30
(20*R/S*) -taraxastanediol	20	A	HSQC, 1D	31
multiflorenol	21	A	HSQC, 1D	32

* GC-MS evidence using Method A, Chapter 2.



Scheme 5.2. Mechanism of the formation of *L. sativa* mixed amyrin synthase products.

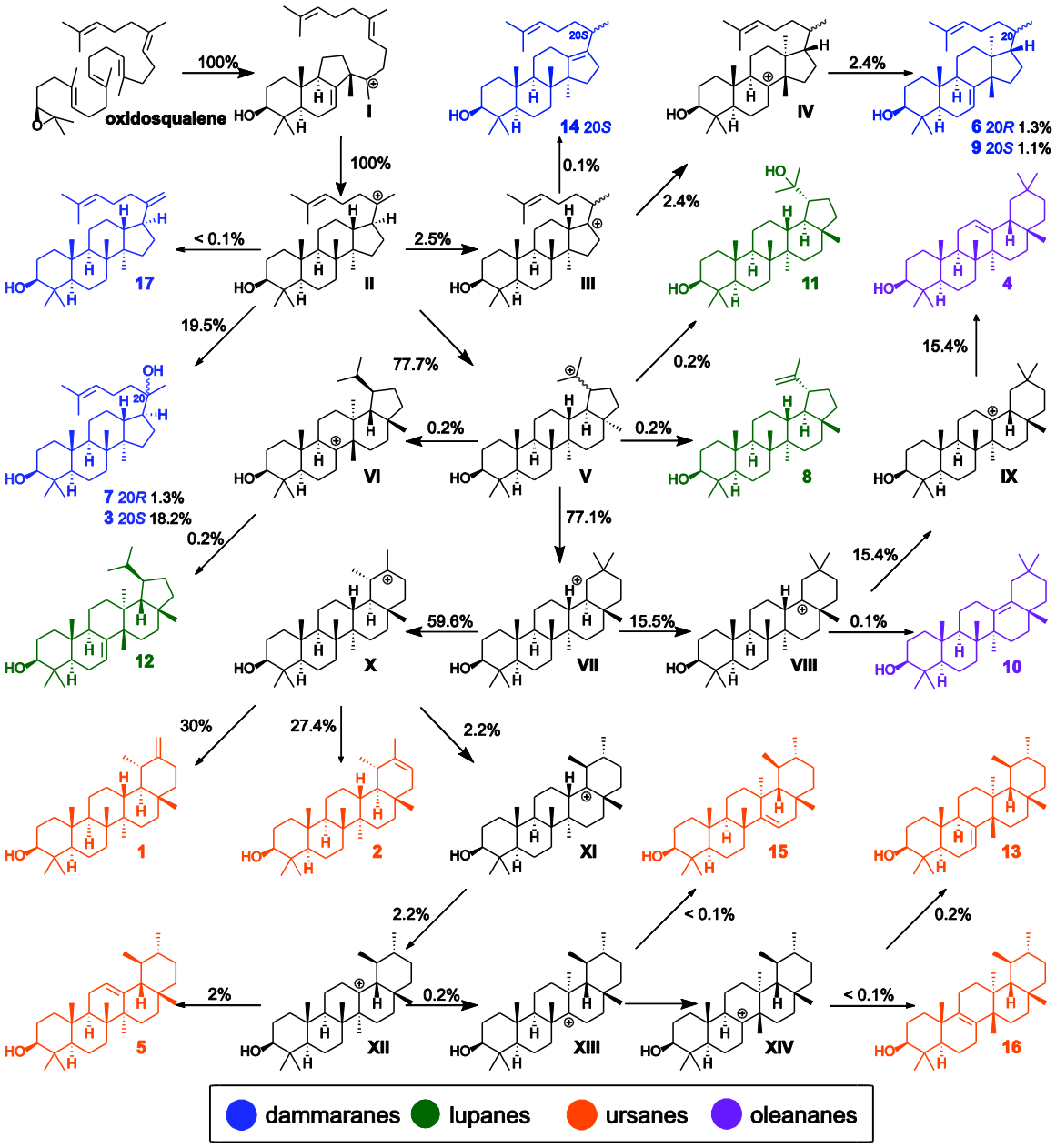
5.3.2.3 *In vitro* Product Profile of Mixed Taraxasterol Synthase

L. serriola mixed taraxasterol synthase (BQ997802) was cloned by Gia Fazio,[1] but the product profile has not been previously reported. This enzyme makes roughly a 1:1 ratio of taraxasterol to ψ-taraxasterol and at least 15 other minor products. The mechanism for this enzyme and product profile ratios are presented in Table 5.9 and Scheme 5.3. Product profiles were determined as described in Chapter 3. Ratios for the triterpene-monoalcohols and diols were determined using an HSQC spectrum of RXY6[pGCF16.9] Fractions 22-32.

Table 5.9. Table summary of analytical evidence for *L. serriola* mixed taraxasterol synthase products identified in the 2 L *in vitro* experiment.

Product	Product #	GC-MS*	NMR Evidence	Reference
taraxasterol	1	A	HSQC, 1D	20
ψ-taraxasterol	2	A	HSQC, 1D	20
(20*S*)-dammarenediol	3	A	HSQC, 1D	21
β-amyrin	4	A	HSQC, 1D	18
α-amyrin	5	A	HSQC, 1D	17
butyrospermol	6	A	HSQC, 1D	19
(20*R*)-dammarenediol	7	A	HSQC, 1D	21
lupeol	8	A	HSQC, 1D	22
tirucalla-7,24-dien-3β-ol	9	A	HSQC, 1D	14
δ-amyrin	10	A	HSQC, 1D	23
lupane-3β,20-diol	11	A	HSQC, 1D	30
nematocyphol	12	A	HSQC, 1D	25
bauerenol	13	A	HSQC, 1D	27
isotirucallol	14	A	HSQC, 1D	15
isoursenol	15	A	HSQC, 1D	24
isobauerenol	16	A	HSQC, 1D	26
dammara-20,24-dienol	17	A	HSQC, 1D	19

* GC-MS evidence using Method A, Chapter 2.



Scheme 5.3. Mechanism of the formation of *L. serriola* mixed taraxasterol synthase products.

5.3.2.4 *In vitro* Product Profile of Lupeol Synthase

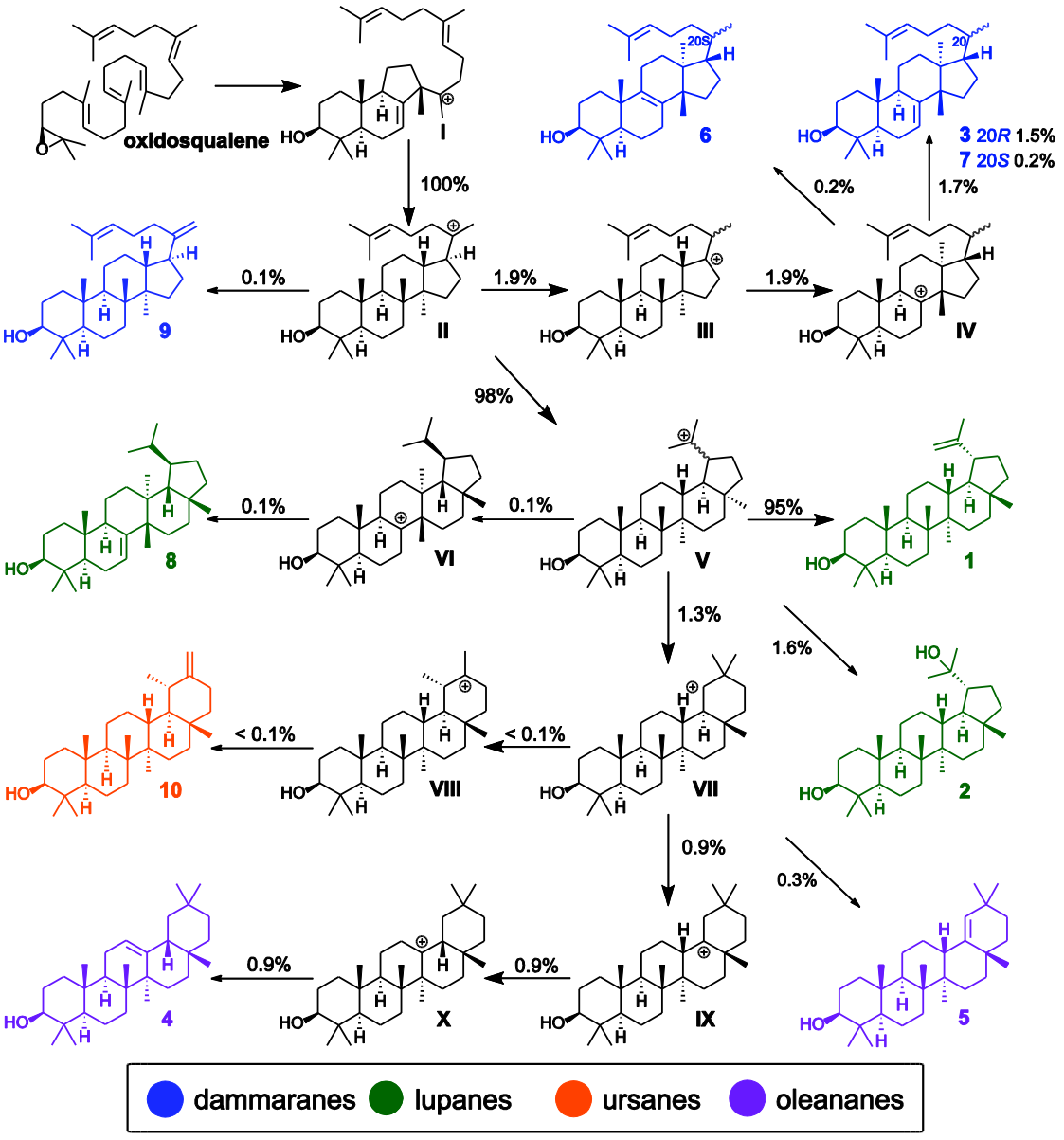
L. serriola lupeol synthase (BU003435) is 92% identical to the *Taraxacum* lupeol synthase. Gia Fazio[1] was the first one to clone this enzyme and reported its product profile. She reported 6 compounds including lupeol (93%), β-amyrin (4.8%), butyrospermol (0.8%), and taraxasterol (0.1%). My new analysis of the product profile revealed a total of ten compounds including germanicol, lupane-3β,20-diol, tirucalla-8,24-dien-3β-ol, tirucalla-7,24-dien-3β-ol and nematocyphol.

The product profile ratios and mechanism of formation are presented in Table 5.10 and Scheme 5.4. Product profiles were determined as described in Chapter 3. Ratios for the triterpene monoalcohols and diols were determined using the HSQC spectrum of RXY6[pGCF17.6] Fractions 21-32.

Table 5.10. Table summary of analytical evidence for *L. serriola* lupeol synthase products identified in the 2 L *in vitro* experiment.

Product	Product #	GC-MS*	NMR Evidence	Ref
lupeol	1	A	HSQC, 1D	22
lupane-3β,20-diol	2	A	HSQC, 1D	30
butyrospermol	3	A	HSQC, 1D	19
β-amyrin	4	A	HSQC, 1D	18
germanicol	5	A	HSQC, 1D	29
tirucalla-8,24-dien-3β-ol	6	A	HSQC, 1D	13
tirucalla-7,24-dien-3β-ol	7	A	HSQC, 1D	14
nematocyphol	8	A	HSQC, 1D	25
dammara-20,24-dienol	9	A	HSQC, 1D	19
taraxasterol	10	A	HSQC, 1D	20

* GC-MS evidence using Method A, Chapter 2.

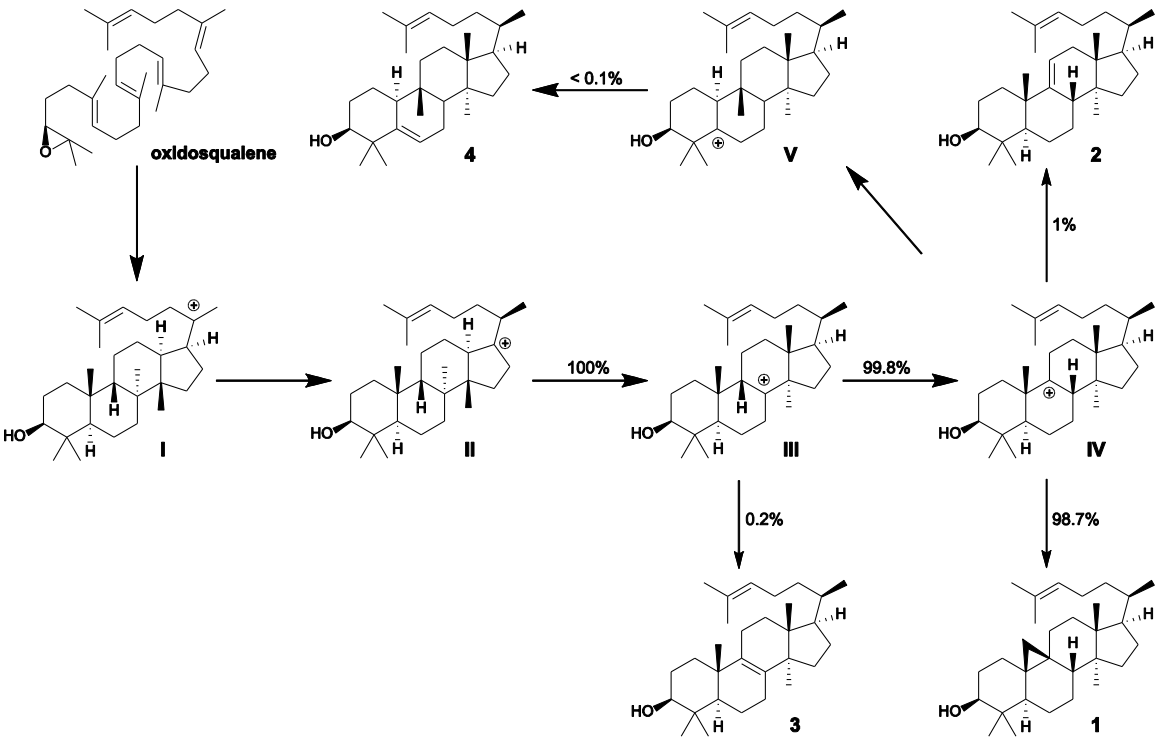


Scheme 5.4. Mechanism of the formation of *L. serriola* lupeol synthase products.

5.3.2.5 *In vitro* Product Profile of Cycloartenol Synthase

L. serriola cycloartenol synthase (BU008606) was first characterized by Gia Fazio.[1] Within the product profile described in her thesis, she reported as major product cycloartenol (87.8%) and parkeol (12.2%). Although this enzyme is in fact a cycloartenol synthase, the product profile reported previously was inaccurate. Table 5.11 and Scheme 5.5 show that cycloartenol accounts for 98.7% of the total of triterpene alcohols that this enzyme can make, followed by 1% parkeol, 0.2% lanosterol and lastly <0.1% curcurbitadienol.

Product profiles were determined as described in Chapter 3. Ratios for the triterpene-monoalcohols were determined using HSQC spectrum of RXY6[pGCF19.6] Fractions 20-33.



Scheme 5.5. Mechanism of the formation of *L. serriola* cycloartenol synthase products.

Table 5.11. Summary of analytical evidence for *L. serriola* cycloartenol synthase products identified in the 2 L *in vitro* experiment.

Product	Product #	GC-MS*	NMR Evidence	Ref
cycloartenol	1	A	HSQC, 1D	33
parkeol	2	A	HSQC, 1D	34
lanosterol	3	A	HSQC, 1D	35
curcurbitadienol	4	A	HSQC, 1D	36

* GC-MS evidence using Method A, Chapter 2.

5.3.2.6 *Lactuca* Plant Extracts Analysis and Curcurbitadienol from Watermelon Seeds Extracts

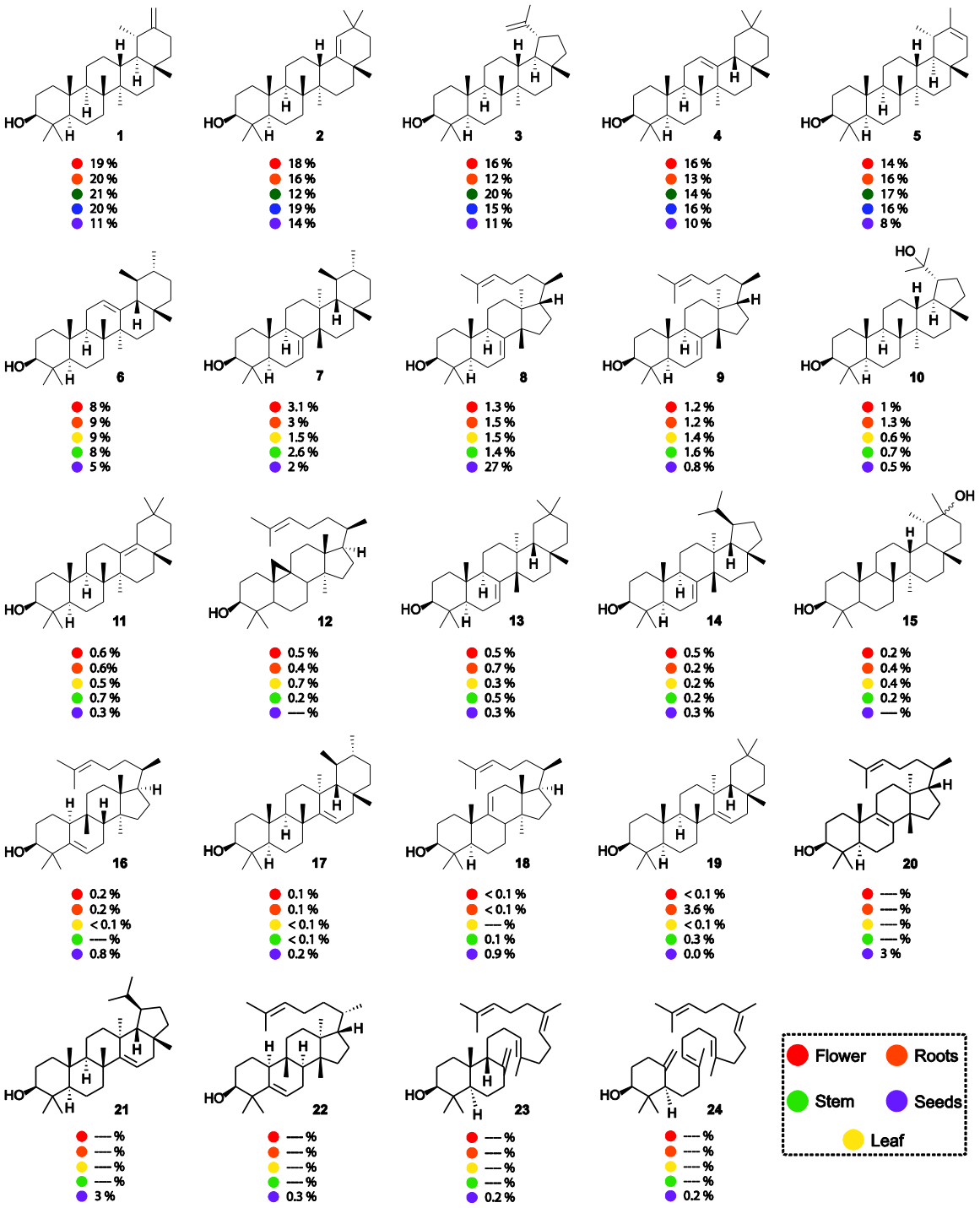
Table 5.12 and Scheme 5.6 have a detailed summary of the triterpene alcohols found in *L. sativa*. Figure 5.9 gives details for the GC-MS and NMR information used to characterize curcurbitadienol that helped to identify this compound as a minor product of *L. serriola* cycloartenol synthase (BU008606).

Table 5.12. Table summaries of names, numbering and analytical evidence for compounds found in lettuce plant extracts.

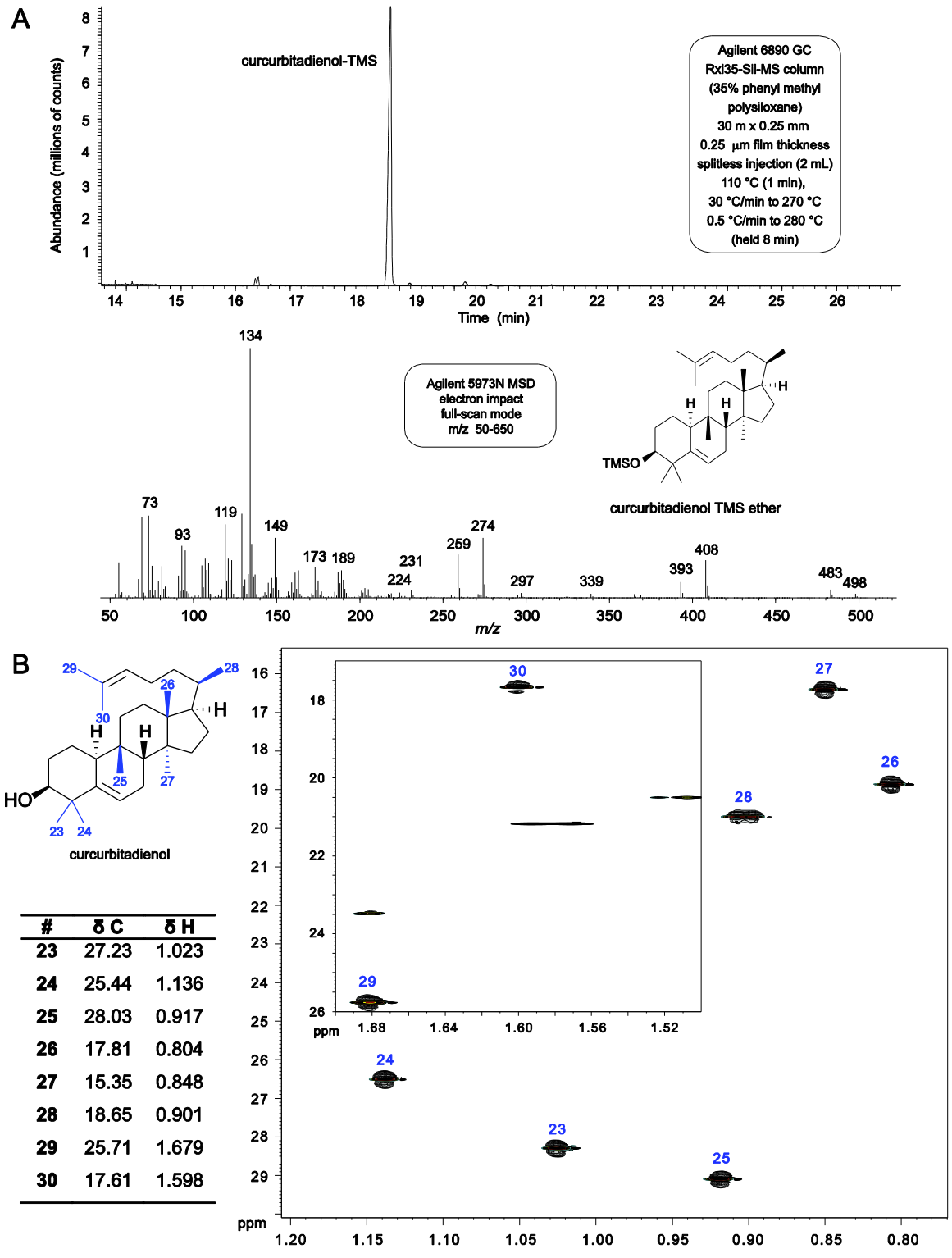
Product	Product #	GC-MS*	NMR Evidence	Ref
taraxasterol	1	A	HSQC	20
germanicol	2	A	HSQC, 1D	29
lupeol	3	A	HSQC, 1D	22
β-amyrin	4	A	HSQC, 1D	18
ψ-taraxasterol	5	A	HSQC, 1D	20
α-amyrin	6	A	HSQC, 1D	17
bauerenol	7	A	HSQC, 1D	27
butyrospermol	8	A	HSQC, 1D	19
tirucalla-7,24-dien-3β-ol	9	A	HSQC, 1D	14
lupane-3β,20-diol	10	A	HSQC, 1D	30
δ-amyrin	11	A	HSQC, 1D	23

cycloartenol	12	A	HSQC, 1D	33
multiflorenol	13	A	HSQC, 1D	32
nematocyphol	14	A	HSQC, 1D	25
(20*R/S*) -taraxastanediol	15	A	HSQC, 1D	31
cucurbitadielnol	16	A	HSQC, 1D	36
isoursenol	17	A	HSQC, 1D	24
parkeol	18	A	HSQC, 1D	34
taraxerol	19	A	HSQC, 1D	28
euphol	20	A	HSQC, 1D	37
tarolupenol	21	A	HSQC, 1D	38
boeticol	22	A	HSQC, 1D	39
Δ8(26)-polypodatetraenol-9βH	23	A	HSQC, 1D	40
achilleol A	24	A	HSQC, 1D	41

* GC-MS evidence using Method A, Chapter 2.



Scheme 5.6. Summary of the products found in the plant extracts of the various parts of lettuce.



#	δ C	δ H
23	27.23	1.023
24	25.44	1.136
25	28.03	0.917
26	17.81	0.804
27	15.35	0.848
28	18.65	0.901
29	25.71	1.679
30	17.61	1.598

Figure 5.11 Analytical evidence of curcurbitadienol in watermelon seeds.

5.4 Discussion

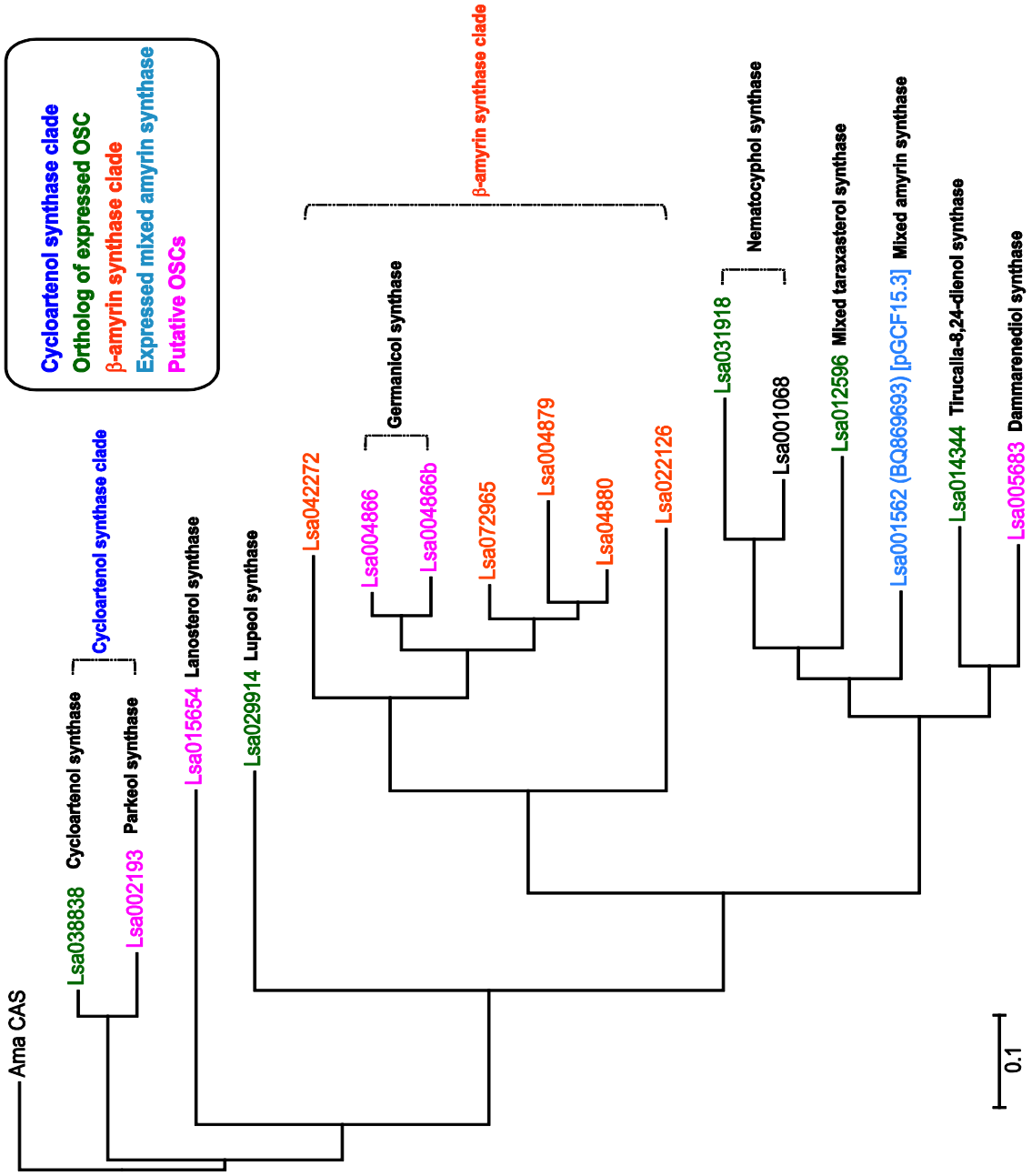


Figure 5.12. Molecular phylogenetic tree of *Lactuca sativa* OSCs found by manual splicing of the genome. Genes within the β-amyrin, germanicol synthase clade and dammarenediol synthase are considered putative. The tree was constructed by maximum likelihood using MEGA version 6.[11] Multiple sequence alignments were generated in MegAlign from amino acid sequences with Clustal W Method using default parameters.

5.4.1 CAS Clade

L. sativa seems to have two putative cycloartenol synthases (Figure 5.8, Figure 5.12), Lsa038838 and Lsa002193. Lsa038838 is not only 99% identical to *L. serriola* cycloartenol synthase (BU008606) but the putative active site sequence alignment is 100% conserved, which means that this is the possible ortholog of the cycloartenol synthase studied in this chapter.

On the other hand, Lsa002193 is only 90% similar to *L. serriola* CAS1 and the active site residues show two substitutions (Leu98Tyr and Leu702Ile) with respect to *L. serriola* cycloartenol synthase. Both amino acids are located near the B ring, where the C-9 cationic intermediate could be potentially deprotonated to cycloartenol, parkeol, lanosterol or curcurbitadienol.

In an article published in May of 2015, Kushiro and co-workers[43] described how a single point mutation (Leu125Tyr) of curcurbitadienol synthase (CBS) from *Cucurbita pepo*, converted the cyclase into a parkeol synthase. The Leu to Tyr mutation seems to prevent the formation of the cyclopropyl ring characteristic of cycloartenol due to possible steric hindrance from the large amino acid side chain of Tyr; it also seems to be involved in the prevention of further rearrangement to yield curcurbitadienol, instead it prefers deprotonation to yield the Δ9(11) double bond characteristic of parkeol.

Alignment of curcurbitadienol synthase from *Cucurbita pepo* and Lsa002193 from *Lactuca sativa* revealed that Leu98Tyr corresponds to Leu125Tyr from Takase *et. al.*[43], thus showing that Lsa002193 could be a putative parkeol synthase from *L. sativa*. Further evidence for the presence of a putative parkeol synthase among the OSCs in *L. sativa* is provided by the unusual amount of parkeol in the lettuce plant extracts (Scheme 5.6). Parkeol is one of the minor products of the cycloartenol synthase (BU008606) from *L. serriola*, however the amount of parkeol made by this enzyme does not account for the amounts of parkeol that are reported in the different lettuce plant extracts. Parkeol is often isolated as a minor cyclization product in cycloartenol and lanosterol synthases; [16] however, sea cucumbers further metabolize this compound to saponins.[44] Saponins are known to confer protection against microbes and other predators.[45] Even though the role of parkeol is unknown in lettuce, it may be involved in plant defense against external factors.

5.4.2 Lanosterol Synthase

Lsa015654 is an OSC from *L. sativa* that could encode for a putative lanosterol synthase. The analysis of the amino acids of the active site, and the overall sequence identity with respect to *A. thaliana* LSS1[46] and *A. lyrata* LSS1 revealed that Lsa015654 has an almost identical active site alignment, except for position 233 in where Lsa015654 has a Thr instead of Cys characteristic of the *Arabidopsis* genus. The overall amino acid sequence identity when compared to two lanosterol synthases from the rosids clade (*Aly*LSS and *Ath*LSS) is 66%. Although this is low identity in comparison with other pairs of orthologs in this work, enzymes like lanosterol synthases across species have overall low sequence identity and still make lanosterol as the major product.[16]

Lanosterol biosynthesis in plants seems to be part of the biosynthesis of polar metabolites that are involved in signaling and defense.[48] In *Euphorbia* species, lanosterol and lanosterol oligosaccharides have been detected in low amounts; it is speculated that the role of lanosterol synthase in plants is related to the biosynthesis of phytosterols, which provide assistance in plant growth and development.[49]

5.4.3 β-amyrin Synthase Clade and Putative Germanicol Synthase

The phylogenetic tree in Figure 5.12 groups seven putative OSCs from *Lactuca sativa*. Due to the high sequence identity of these OSCs to β-amyrin synthases from asterids, this clade is called: β-amyrin clade.

Out of these seven genes, Lsa072965 is the best candidate to be a β-amyrin synthase since it shares 90% sequence identity to β-amyrin synthases from *Olea europea*, *Taraxacum officinale* and *Artemisia annua*. [50] On the other hand, Lsa042272, Lsa004879, Lsa04880 and Lsa022126 identities are only in the high 60s to low 70s, making it more difficult to hypothesize which type of cyclases this group could encode.

Results from plant extracts show that butyrospermol is a compound well distributed among lettuce plant parts; however it is present in high amounts (27% of triterpenes) in seeds (as shown in Scheme 5.6 and Table 5.12), suggesting a possible butyrospermol synthase among the OSCs whose product profiles have not been determined.

Lastly, Lsa004866 and Lsa004866b are genes that seem to have arisen from gene duplication within the genome (paralogs). DNA analysis and blast searching revealed that these genes are 92% identical, and their active site sequences differ only at positions 236 and 337. They are within the clade that has the lowest amount of characterized cyclases (orthologs or otherwise) and they are the pair that has highest sequence identity to the known germanicol synthase from *Rhizophora stylosa* [51] (75% for Lsa004866b and 77% for Lsa004866). Further evidence that *Lactuca sativa* might have one or two germanicol synthases is that germanicol is the second major compound (besides □-amyrin) accumulating widely in different parts of the lettuce plant (Scheme 5.7).

5.4.4 Putative Dammarenediol Synthase and Nematocyphol Synthase Orthology

Lsa001562 (BQ869693) is a mixed amyrin synthase whose product profile and phylogeny will be discussed in more depth later on. The rest of the cyclases in this clade have been expressed and characterized by Fazio and by myself in the closest organism *L. serriola* (Figure 5.14 and Figure 5.15).

Each pair of orthologs shows a 99% sequence identity, as well as 100% active site sequence identity. This high level of homology could be explained by the fact that cultivated *Lactuca* species are highly inbred,[52] in order to retain certain traits like tastes and pathogen resistance from *L. serriola*.[53]

The only two genes from this clade that are not characterized in either *L. sativa* or *L. serriola* are Lsa001068 and Lsa005683.

Lsa031918 and Lsa001068 are grouped as putative nematocyphol synthases (Figure 5.12, Figure 5.13), Lsa031918 has higher sequence homology (95%) than Lsa001068 (83%) to the *L. perennis* nematocyphol synthase from Chapter 4. It is more likely that Lsa031918 is the nematocyphol synthase ortholog in *L. sativa*.

Lsa005683 might be a dammarenediol synthase since its closest homolog of Lsa005683 within the asterids clade is *Panax notoginseng* dammarenediol synthase (77% sequence indentity).[54]

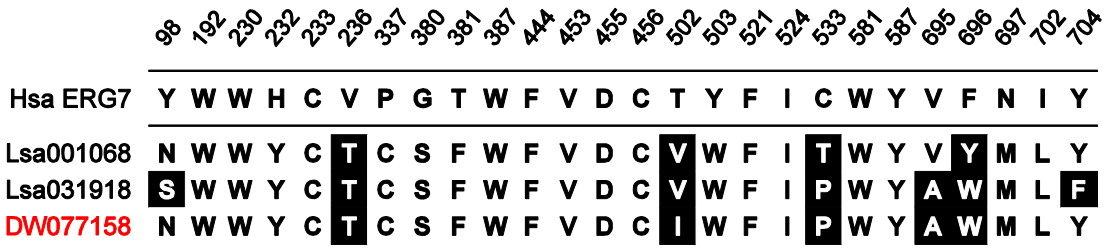


Figure 5.13. Comparison of active site residues of NEM putative orthologs.

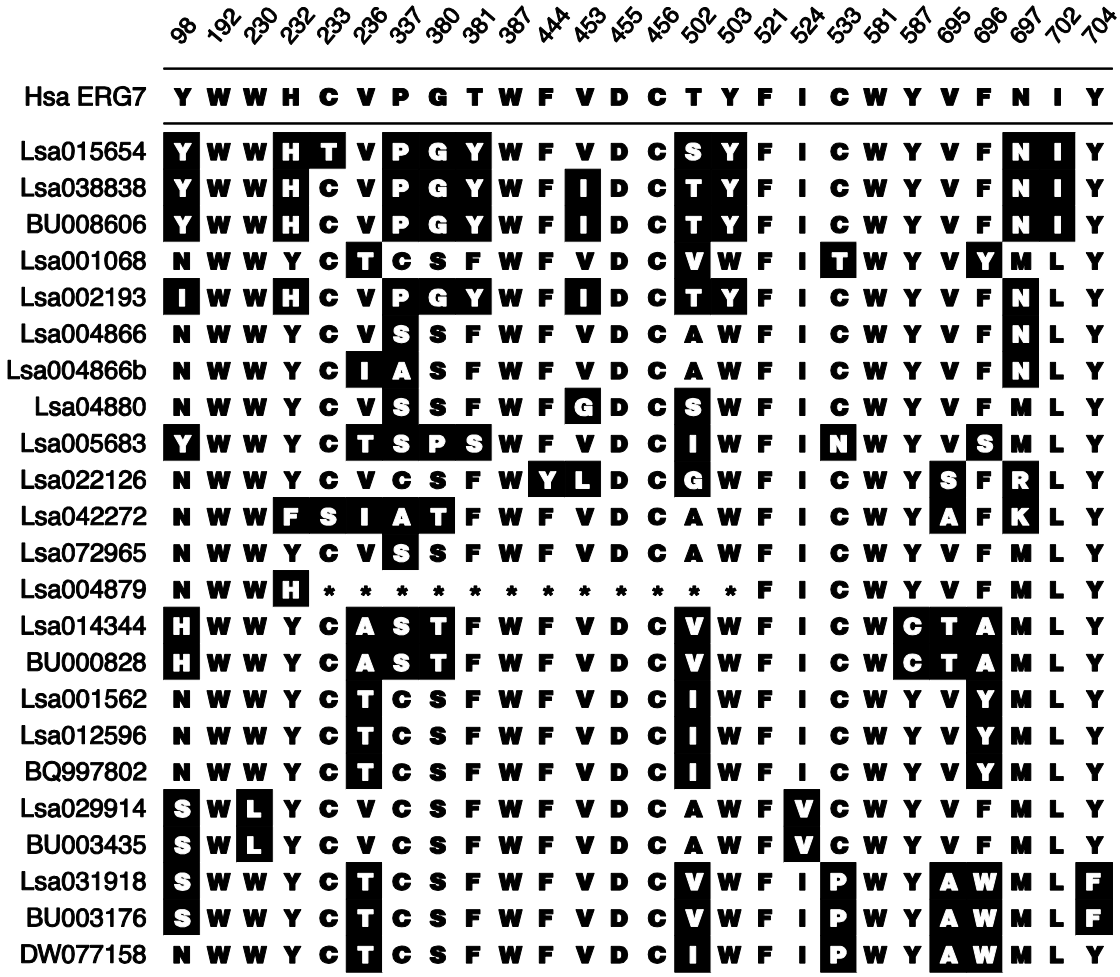


Figure 5.14. Comparison of active-site residues in human LSS[42] and corresponding residues of *Lactuca* OSCs from *L. sativa*, *L. serriola* and *L. perennis.*

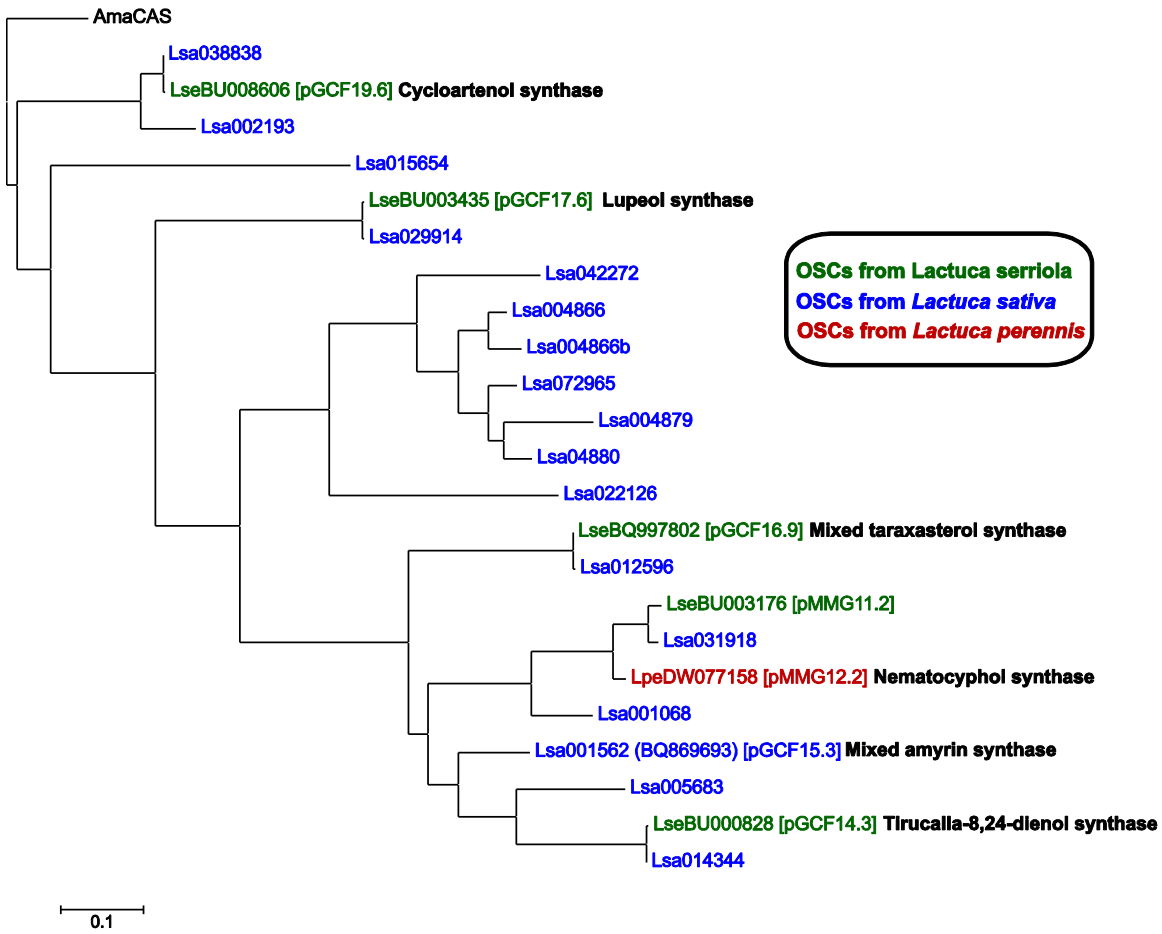


Figure 5.15. Molecular phylogenetic tree of *Lactuca* OSCs. Expressed OSCs from the *Lactuca* family (*L.sativa* or *L. serriola*) are the ones preceded by the name in black. The tree was constructed by maximum likelihood using MEGA version 6.[11] Multiple sequence alignments were generated in MegAlign from amino acid sequences with Clustal W Method using default parameters.

5.4.5 Tirucalla-8,24-dien-3β-ol Synthase from *Lactuca serriola*

Tirucalla-8,24-dien-3β-ol synthase from *Lactuca serriola* is 99.7% identical to its ortholog in *Lactuca sativa*, and the active site amino acids are the same. Phylogenetic comparison of this pair of these orthologs indicates that *L. sativa* also has a tirucalla-8,24-dien-3β-ol synthase, with a very similar product profile.

Within the asterids clade, this is the first ever described tirucalla-8,24-dien-3β-ol synthase. Early work from Jazmin Godoy and Caroline McNeil supports the product profile described in the results section of Chapter 5. This OSC was the only enzyme that did not performed well during *in vitro* expression, for this reason only preliminary *in vivo* results are presented in this work.

The product profile for this enzyme includes tirucalla-8,24-dien-3β-ol **1** (70%), tirucalla-7,24-dien-3β-ol **2** (15%), isotirucallol **3** (5%), (20*S)*-dammara-12,24-dienol **4** (2%) and a mixture of unknown minor products that account for the remaning 8% of total products. All the products made by this enzyme are 20*S* tetracyclic epimers that arise from **II** with the side chain *syn* to C-16 when 17α-H migrates or side chain anti to C-16 when 17β-H migrates.[55]

The other known characterized enzyme that makes predominantly 20*S* dammaranes is *Ath*LUP5, whose major product is tirucalla-7,24-dien-3β-ol.[16,56] Classification of product accuracy (Figure 5.16) for tirucalla-8,24-dien-3β-ol synthase places this enzyme in the moderately accurate category (P_1/P_2: 4.5 / $P_1/\sum P_i$: 0.7), whereas *A. thaliana* LUP5 is a non-product-specific cyclase (P_1/P_2: 1.1/ $P_1/\sum P_i$: 0.36) or multifunctional cyclase.[16] The sequence identity between these two enzymes from distant clades is only 53%, and several active site residues are different. Analysis of the plant extracts does not show any tirucalla-8,24-dien-3β-ol present in the plant, so the specific role of this cyclase in *Lactuca* is unknown.

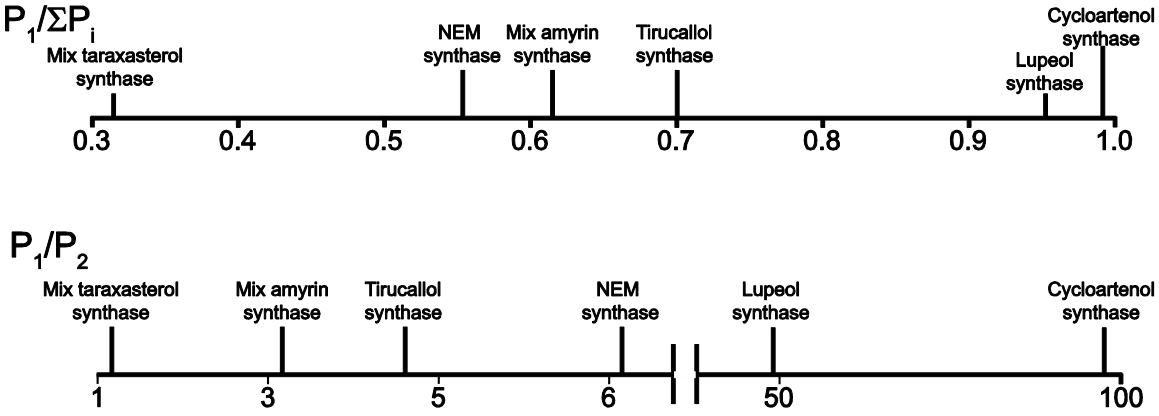


Figure 5.16. Classification of the characterize OSC from *Lactuca* genus by its accuracy. $P_1/\sum P_i$ is the ratio of the major product to the total amount of products; P_1/P_2 is the ratio of the major product to the second most abundant product.

5.4.6 Mixed Amyrin Synthase from *Lactuca sativa*

This mixed amyrin synthase (BQ869693) was previously reported by Gia Fazio, and it is the only cyclase from the *L. sativa* family for which the product profile has been characterized. While Gia Fazio[1] reported a 3:1 ratio β-amyrin to α-amyrin and seven other minor products, a decade of improvement in our analytical techniques provided a more detailed product profile: 3:1 ratio of α-amyrin to β-amyrin (~80% of total products), with minor compounds that include dammaranes, lupanes, ursanes and oleanes.

Cyclization starts by converting OS into the dammarenyl cationic intermediate **II**. Cation quenching by water at C20 yields (20*S*)-dammarenediol **6** (2.3%), (20*R*)-dammarenediol **9** (0.7%). Direct deprotonation of **II** without rearrangement gives dammara-20,24-dienol **3** (5.3%). The remaining 91.7% of the material undergoes a series of hydrogen and methyl 1,2-shifts (**III** and **IV**) followed by hydride elimination to yield isotirucallol **16** (0.1%), tirucalla-7,24-dien-3β-ol **5** (2.8%) and butyrospermol **7** (0.8%).

Then the remaining 88% of the material undergoes D-ring expansion followed by E-ring closure to give the lupanyl cation **V**. Direct deprotonation of **V** without rearrangement provides lupeol **10** (0.6%), and water quenching of the C20 cation forms lupane-3β,20-diol **18** (<0.1%). After a series of 1,2-shifts, deprotonation from **V** generates

nematocyphol **13** (0.2%). The material that is lost en route to the formation of lupane skeletons is less than 1% of the total.

This enzyme favors formation of oleanes and ursanes; 87.1% of the material will convert from **V** to the cationic intermediate **VII** by E-ring expansion. Direct deprotonation of germanicyl cation **VII** yields germanicol **18** (<1%), and 19.7% of the material undergoes a series of 1,2-shifts followed by deprotonation to generate δ-amyrin **11** (0.6%), multiflorenol **21** (<0.1%), taraxerol **17** (<0.1%) and β-amyrin **2** (19.1%).

En route to the formation of ursane type compounds, 5.1% of the material will rearrange from the germanicyl cation **VII** to **XII**. Proton elimination from the newly formed taraxastyl cation **XII** yields taraxasterol **8**, ψ–taraxasterol **4** and 3β(20)*R/S*-taraxastanediol **20**, which are made in 0.7%, 4.2% and <1% amounts, respectively.

On the other hand, when **VII** rearranges to the ursanyl cation **XIII**, a series of 1,2-shift and proton eliminations give the major product α-amyrin **1** (61.7%), isoursenol **12** (0.2%), baurenol **15** (0.1%) and isobauerenol **14** (0.1%). The ursane type compounds that derived from **XIII,** account for 62.1% of the total products. En route to **1**, little material is lost to due to formation of minor products, which suggests that once the enzyme makes the taraxasteryl cation, it is good at controlling formation of **1** without a significant lost to minor "mistakes".

A refined product profile of this enzyme allows placing it in the accuracy continuum shown in Figure 5.16. The values for accuracy classification are P_1/P_2: 3.22 and $P_1/\sum P_i$: 0.61; classification following Paul Bodager[16] places this enzyme close to the moderately product-specific category.

Analysis of extracts of different parts of lettuce (Scheme 5.6) shows that in *L. sativa*, β-amyrin is present in a higher level than α-amyrin, approximately in a 2:1 ratio. Clearly other cyclases are responsible for the abundance of β-amyrin in lettuce. In the results section it was suggested that lettuce could have a cyclase that makes β-amyrin as a major product; this would explain in part the ratio discrepancy between enzymatic product profile and plant extracts results or α-amyrin could be metabolize by the plant.

Even though this might be the first example of an asterid OSC that makes predominantly amyrin products, there is only one other characterized mixed amyrin synthase from the rosids clade. *Pisum sativum* cyclase makes a 3:2 ratio of α-amyrin to β-

amyrin,[57] accompanied by several other minor byproducts. However, sequence identity with our cyclase is only 60%. This cyclase is a nice example of parallel evolution of mixed amyrin synthases in distant clades.

5.4.7 Mixed Taraxasterol Synthase from *Lactuca serriola*

Taraxasterol and ψ-taraxasterol[20] are known compounds that are present as minor products in *A. thaliana*. Mixed taraxasterol synthase (BQ997802) from *L. serriola* is the first example of an enzyme that synthesizes this compound in a 1:1 ratio and as major products. The complete profile includes 15 minor products. The phylogenetic tree shows that this cyclase shares a common ancestor with NEM, mixed amyrin synthase and tirucallol synthase; it is grouped in a separate branch along with its homolog in *L. sativa* (Figure 5.12 and 5.15). Mixed taraxasterol synthase from *L. sativa* and the expressed clone from *L. serriola* are 99.6% identical and have the same amino acid sequence for the active site.

Plant extracts show that both compounds are predominant in all the parts of the lettuce plant and the ratios are quite close to the ones found in this cyclase, discrepancies could be accounted for by contributions of minor amounts of taraxasterols from other OSCs.

The mechanism of formation of the product profile of this multifunctional cyclase (Table 5.9 and Scheme 5.3) starts by cyclization of oxidosqualene to the 6-6-6-5 dammarenyl cation **II**. Twenty percent of the material is lost to the formation of (20*S*)-dammarenediol **3** (19.5%), (20*R*)-dammarenediol **7** (1.3%) and dammara-20,24-dienol **17** (<0.1%); these pathways involve addition of a hydroxyl group at C20 or direct deprotonation of **II** without rearrangement.

Around 2.5% of the total is converted to tetracyclic products following a series of hydride and methyl 1,2-shifts via **III** and **IV** to form isotirucallol (**14)**, tirucalla-7,24-dien-3β-ol (**9)** and butyrospermol (**6)** at levels of 0.1%, 1.3% and 1.1%. The remaining 77.7% goes into D-ring expansion followed by E-ring closure and gives the lupanyl cation **V**.

En route to the formation of taraxastane, ursane and oleane skeletons, only 0.6% of the material is converted to lupanes via rearrangement or water addition to the C20 of **V**. The results are lupeol **8** (0.2%), lupane-3β,20-diol **11** (0.2%) and nematocyphol **12** (0.2%).

Germanicyl cation **VII** (77.1%) arises from E-ring expansion of **V**; less than one third of the material rearranges to form the oleanyl cation **VIII**, the rest forming **X** (59.4%) in route to the major product **1** or to **VIII**, which will rearrange to β-amyrin **4** (15.4%) and δ-amyrin **10** (0.1%).

Deprotonation of the taraxastyl cation **X** (59.4%) yields taraxasterol **1** (30%) and ψ–taraxasterol **2** (27.4%). A small portion of **X** (2.2%) undergoes three 1,2-hydride shifts via **XI** to form the ursanyl cation **XII**, which is either directly deprotonated to α-amyrin **5** (2%) or undergoes further rearrangement to isoursenol 1**5** (<0.1%), baurenol **13** (0.2%) and isobauerenol **16** (0.1%). P_1/P_2 and $P_1/\sum P_i$ values place this enzyme in the non- product-specific category. The four structurally diverse major products, Taraxasterol, ψ-taraxasterol, (20*S*)-dammarenediol, and β-amyrin (30:27:20:15 ratio) comprise 92% of total products. The diverse structures might provide some kind of shotgun defense; taraxasterol has anti-inflamatory properties in human cell culture.[58]

5.4.8 Lupeol Synthase from *Lactuca serriola*

Lupeol synthase from *L. serriola* was previously described by Gia Fazio[1], who identified six products and four unknowns. This work identifies the minor unknowns and also analyzes the phylogenetic relationship of this cyclase within the *L. sativa* genome. This OSC is 99.5% identical to the *L. sativa* putative lupeol synthase ortholog. Alignment of the amino acids that are believed to be involved in catalysis shows that their active site is 100% identical. This high level of homology suggests very similar product profiles.

Lupeol synthase (BU003435) falls within the category of accurate lupeol synthases (P_1/P_2: 47.3 $P_1/\sum P_i$: 0.95) and is 92% identical to another accurate lupeol synthase[59] from *Taraxacum officinale* (Asteraceae family). Other accurate lupeol synthases in asterids isolated from *Olea europaea*[59] and *Betula platyphylla*[60] also show ≥76% identity with BU003435. No OSC in *A. thaliana* makes more than ca. 20% lupeol, although the literature[55] might suggest higher amounts.

Lupeol has been shown to be the dominant component of the epicuticular wax of the castor bean *Ricinus communis*.[61] In animal tests lupeol shows anti-cancer[62] and anti-malarial[63] capabilities as well as activity as a hepatoprotectant,[64] and oxidized forms of

lupeol can have potent bioactivities. The role of lupeol or its metabolites in *Lactuca* remains to be explored.

As shown in Scheme 5.4, BU003435 cyclizes oxidosqualene to the dammarenyl cation **II**, 2% of which will form tetracycles: dammara-20,24-dienol **9** (0.1%), tirucalla-8,24-dien-3β-ol **6** (0.2%), tirucalla-7,24-dien-3β-ol **7** (0.2%) and butyrospermol (1.5%) **3**. The remaining 98% undergoes D-ring expansion and E-ring closure to give the lupanyl cation **V**, which is mainly deprotonated to lupeol **1** (95%). Rearrangement of **V** by 1,2-shifts gives nematocyphol **7** (0.1%), and water addition to **V** gives lupane-3β,20-diol **2** (1.6%). The remainder of **V** (<2%) undergoes E-ring expansion and 1,2-shifts to give mainly β-amyrin **4** (0.9%), taraxasterol **10** (<0.1%) and germanicol **5** (0.3%).

5.4.9 Cycloartenol Synthase from *Lactuca serriola*

The last cyclase to be discussed in this work is involved in plant primary metabolism. As shown in Scheme 5.5, *L. serriola* cycloartenol synthase (BU008606) cyclizes oxidosqualene to cycloartenol **1** (98.7%), and other 3 minor products originating from the protostery cation **I**.

Cation **I** arises from oxidosqualene cyclization via chair-boat-chair intermediates, instead of the chair-chair-chair cations of secondary metabolism). Then a cascade of 1,2-hydride and methyl shifts generates the lanosteryl cation **III**, from which C9 deprotonation gives lanosterol **3** (0.2%). Another hydride shift gives cation **IV**, which is deprotonated at C11 to parkeol **2** (1%) or cyclized with deprotonation to cycloartenol **1** (98.7%). A trace of **IV** undergoes further 1,2-shifts to cation **V**, which is deprotonated to curcurbitadienol **4** (<0.1%).

L. serriola cycloartenol synthase is 99.7% identical to its *L. sativa* ortholog and they share the same active site alignment. The cycloartenol synthase from *A. thaliana*, is only ~80% identical to the *Lactuca* cycloartenol synthases, however the active site is 100% identical.

Paul Bodager,[16] performed an extensive analysis on OSCs from primary metabolism, and while he found as a minor product curcurbitadienol in human lanosterol synthase and cow lanosterol synthase. He did not find any curcurbitadienol from *A.*

thaliana cycloartenol synthase, the only plant OSC that he analyzed but did identify curcurbitadienol as a minor product of other OSCs. He also showed that primary metabolism in plant, animals and fungi can tolerate OSCs that make up to around 1% total byproducts, while maintaining efficient production of essential sterol precursors.

5.5 Genome Mining to Elucidate the Triterpene Biosynthetic Capability of the *Lactuca* Clade

The availability of the genome of an organism presents the opportunity to obtain a complete picture of the genes involved in secondary metabolism. In the study of OSCs, the accessibility of other plant genomes opens up the possibility to explore differences and similarities in triterpene biosynthesis across species. *A. thaliana* has given a general insight of OSCs evolution and has provided the foundation for detailed characterization of plant genes.[65]

The analysis of OSCs in *Lactuca* was significantly more challenging than the analysis of *A. lyrata* in Chapter 3. Although the experimental analysis was similar to what was done with *Arabidopsis*, the approach had to be modified as *L. sativa* itself might possibly be viewed as the model organism of the Compositae family.[66,67] There was no previous reference work about OSCs in *Lactuca* beyond that carried out by Gia Fazio.

In Chapter 5 it was shown that the *L. sativa* genome encodes 17 OSCs, six of which have characterized orthologs from the closely related organisms *L. serriola* and *L. perennis*. Within the *Lactuca* clade, differences between cyclases (both active site and sequence alignment) are minimal, which implies a higher level of orthology than, for example, the OSCs within the Brasicaceae family (Chapter 3).

Upon sequence analysis, the *Arabidopsis* and *Compositae* clades have only one OSC in common, cycloartenol synthase, which is involved in primary metabolism. The rest of the OSCs appear to be unique to each clade even though most of the triterpene skeletons that have been characterized within the Rosids clade are present in Asterids clade. This is an example of convergent evolution in organisms that are distantly related. They independently evolved oxidosqualene cyclases capable of synthesizing a comparable triterpene product profile.

While enzymes in both *Arabidopsis* and *Lactuca* display the majority of the known triterpene skeletons as part of their product profile, it appears that the chemical solution to common problems is often dissimilar in different plant linages (different sets of OSCs). This idea is supported by the low level of orthology observed among OSCs from Rosids and Asterids.[68] *Lactuca* does have a characteristic set of triterpene compounds (lupane type rearranged products) that have not been detected in *A. thaliana*. Even though the function of these compounds is not yet known, this unique set of secondary metabolites may be related to the mechanism of adaptation and defense that *Lactuca* developed along its evolution pathway.

The analysis of OSCs in *Lactuca* shows that the study of OSCs in distant organisms still depends upon heterologous yeast expression and plant extracts. In spite of the fact that orthology across species seems to be conserved in enzymes involved in primary metabolism, this does not simply translate to secondary metabolites. The prediction of the existence of orthologs for secondary metabolites across species or, in particular, the product profile of OSCs, is more complicated than it was previously thought.

REFERENCES

[1] Fazio, G. *Ph.D. Dissertation*, Rice University **2006**.

[2] Cronquist, A. The Compositae revisited; Brittonia; **1977**, *29*, 137-153.

[3] Agricultural Marketing Resource Center Home Page. http://www.agmrc.org/ (accessed Nov 11, 2015).

[4] Kesseli, R. V.; Michelmore, R. W. *The Compositae: systematically fascinating but specifically neglected. In: Genome Mapping in Plants*; A.H. Paterson ed. R.G. Landes Co.: Georgetown, TX., 1997; pp 179-191.

[5] (a) Kuang, H.; Ochoa, O. E.; Nevo, E.; Michelmore, R. W. *Plant J.* **2006**, *47*, 38-48. (b) Kuang, H.; Woo, S. S.; Meyers, B. C.; Nevo, E.; Michelmore, R. W. *Plant Cell* **2004**, *16*, 2870-2894. (c) Meyers, B. C.; Shen, K. A.; Rohani, P.; Gaut, B. S.; Michelmore, R. W. *Plant Cell.* **1998**, *10*, 1833-1846.

[6] Lavelle, D.; Reyes-Chin-Wo, S.; Song, C.; Kozik, A.; Froenicke, L.; Truco, M. J.; Beitel, C.; Xu, X.; Bicheng, Y.; Michelmore, R. *Sequencing the Genomes of Cultivated Lactuca sativa and its Wild Progenitor, L. serriola*, Plant & Animal Genome XXI Conference, San Diego, CA, January 12-16, 2013.

[7] Lettuce Genome Resource. https://lgr.genomecenter.ucdavis.edu/ (accessed June, 2014).

[8] Truco, M. J.; Reyes-Chin-Wo, S; Lavelle, D.; Kozik, A.; Froenicke, L.; Xu, H.; Michelmore, R. *Lettuce Genome Assembly, Structure and Annotation*, Plant & Animal Genome XXII Conference, San Diego, CA, January 11-15, 2014.

[9] SeqMan NGen®. Version 5.03 DNASTAR. Madison, WI. http://www.dnastar.com/t-nextgen-seqman-ngen.aspx

[10] Geer L. Y.; Marchler-Bauer, A.; Geer, R. C.; Han, L.; He, J.; He, S.; Liu C.; Shi, W.; Bryant, S. H. *Nucleic Acids Res***. 2010**, *38*, D492-D496.

[11] Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. *Mol. Biol Evol*. **2013**, *30*, 2725-2729.

[12] Shan, H.; Wilson, W. K.; Phillips, D. R.; Bartel, B. Matsuda, S. P. T. *Org. Lett*. **2008**, *30*, 1897-1900.

[13] Gonzales, A. G.; Gutierrez, Jerez, F.; Luque Escalona, M. *An. Quim* **1973**, *69*, 921−928.

[14] Itoh, T.; Tamura, T.; Matsumoto, T. *Lipids* **1976**, *11*, 434-441.

[15] Akihisa, T.; Yasukawa, K.; Kimura, Y.; Takase, S.-I.; Yamanouchi, S.; Tamura, T. *Chem. Pharm. Bull.* **1997**, *45*, 2016-2023.

[16] Bodager, P. *Ph.D. Dissertation*, Rice University **2014**.

[17] Chatterjee, A.; Kundu, A. B.; Chakrabortty, T.; Chandrasekharan, S. *J. Chem. Soc. Commun.* **1968**, 418-419.

[18] Bischof, B.; Jeger, O.; Ruzicka, L. *Helv. Chim. Acta* **1949**, *32*, 1911-1921.

[19] Goad, L.J.; Akihisa, T.; Analysis of Sterols; Blackie (Chapman & Hall): London, **1997**.

[20] Ames, T. R.; Beton, J. L.; Bowers, A.; Halsall, T. G.; Jones, E. R. H. *J. Chem. Soc.* **1954**, 1905- 1919.

[21] Mills, J. S. *J. Chem. Soc.* **1956**, 2196-2202.

[22] Ames, T. R.; Halsall, T. G.; Jones, E. R. H. *J. Chem. Soc.* **1951**, 450-457.

[23] Bauer, S.; Schulte, E.; Their, H. P. *Eur Food Res Technol* **2004**, *219*, 487–491.

[24] (a) Laird, W.; Spring, F. S.; Stevenson, R. *J. Chem. Soc.* **1961**, 2638− 2642. (b) Agata, I.; Corey, E. J.; Hortmann, A. G.; Klein, J.; Proskow, S.; Ursprung, J. *J. Org. Chem.* **1965**, 30, 1698−1710. (c) Chivers, H.; Corbett, R. E.; Heng, C. K.; Wilkins, A. L. *J. Chem. Soc. C* **1973**, 1814-1816.

[25] Cao, D.; Su Y. L.; Yang J. S. *Acta Pharmacol. Sin.* **1991**, *27*, 445-451.

[26] Talapatra, S. K.; Sengupta, S.; Talapatra, B. *Tetrahedron Lett.* **1968**, 5963-5968.

[27] (a) Lahey, F. N.; Leeding, M. V. *Proc. Chem. Soc.* **1958**, 2196-2202. (b) Fukuoka, M.; Natori, S. *Chem. Pharm. Bull.* **1972**, *20*, 974-979.

[28] Brooks, C. J. W. *Chem. Ind.* **1953**, 1178.

[29] (a) David, S. *B. Soc. Chim. Fr.* **1949**, 155-160. (b) David, S. *B. Soc. Chim. Fr.* **1950**, 169-172.

[30] (a) Herrera, J. B. R.; Bartel, B.; Wilson, W. K.; Matsuda, S. P. T. *Phytochemistry* **1998**, *49*, 1905-1911. (b) Segura, M. J. R.; Meyer, M. M.; Matsuda, S. P. T. *Org. Lett.* **2000**, *2*, 2257-2259.

[31] Susunaga, G. S.; Siani, A. C.; Pizzolatti, M. G.; Yunes, R. A.; Delle Monache, F. *Fitoterapia* **2001**, *72*, 709-711.

[32] (a) Khastgit, H. N.; Segupta, P. *Chem. Ind.* **1961**, 1077-1078. (b) Segupta, P.; Khastgit, H. N. *Tetrahedron* **1963**, *19*, 123-132.

[33] Bentley, H. R.; Henry, J. A.; Irvine, D. S.; Spring F. S. *J. Chem. Soc.* **1953**, 3673-3678.

[34] Schreiber, K.; Osske, G. *Tetrahedron* **1964**, *20*, 1803-1805.

[35] (a) Curtis, R. G. *Nature* **1952**, *170*, 321-322.

[36] Akihisa, T.; Ghosh, P.; Thakur, S.; Rosenstein, F. U.; Tamura, T.; Matsumoto, T. *J. Jpn. Oil Chem Soc.***1986**, *36*, 1036-1040.

[37] Goad, L.J.; Akihisa, T.; Analysis of Sterols; Blackie (Chapman & Hall): London, **1997**.

[38] Ageta, H.; Shiojima, K.; Masuda, K.; Lin, T. *Tetrahedron Lett.* **1981**, *22*, 2289-2290.

[39] Ferreira, M. J. U.; Asenso, J. R.; Tavares, O. S. *J. Nat. Prod.* **1995**, *78*, 275-279.

[40] Nguyen, L. H. D.; Harrison, L. J. *Phytochemistry* **1998**, *50*, 471-476.

[41] Barrero, A. F.; Alvarez-Manzaneda, E. J.; Alvarez-Manzaneda, R. *Tetrahedron Lett.* **1989**, *30*, 3351-3352.

[42] Thoma, R.; Schulz-Gasch, T.; D'Arcy, B.; Benz, J.; Aebi, J.; Dehmlow, H.; Hennig, M.; Stihle, M.; Ruf, A. Nature **2004**, *432*, 118-122.

[43] Takase, S.; Saga, Y.; Kurihara, N.; Naraki, S.; Kuze, K.; Nakata, G.; Araki, T.; Kushiro, T. *Org. Biomol. Chem.* **2015**, *13*, 7331-7336.

[44] (a) Kerr, R. G.; Chen, Z. *J. Nat. Prod.* **1995**, *58*, 172-176. (b) Makarieva, T. N.; Stonik, V. A.; Kapustina, I. I.; Boguslavsky, V. M.; Dmitrenoik, A. S.; Kalinin, V. I.; Cordeiro, M. L.; Djerassi, C. *Steroids* **1993**, *58*, 508-517. (c) Dhimakaran, D. I.; Lipton, A. P. *SpringerPlus* **2014**, *3* (673).

[45] (a) Osbourn, A. E. *Trends in Plant Science* **1996**, *1*, 4-9. (b) Tava, A.; Odoardi, M. *Adv Exp Med Biol* **1996**, *405*, 97-109.

[46] Kolesnikova, M. D.; Xiong, Q.; Lodeiro, S.; Hua, L.; Matsuda, S. P. T., *Arch. Biochem. Biophys.* **2006**, *447*, 87-95.

[47] Shibuya, M.; Zhang, H.; Endo, A.; Shishikura, K.; Kushiro, T.; Ebizuka, Y. *Eur. J. Biochem.* **1999**, *206*, 302-307.

[48] O'Brien, M.; Chantha, S.; Rahier, A.; Matton, D. P. *Plant Physiol.* **2005**, *139*, 734-749.

[49] Suzuki, M.; Kamide, Y.; Nagata, N.; Seki, H.; Ohyama, K. *Plant J.* **2004**, *37*, 750-761.

[50] (a) Kirby, J.; Romanini, D. W.; Paradise, E. M.; Keasling, J. D. *FEBS J.* **2008**, *275*, 1852-1859. (b) Kushiro, T.; Shibuya, M.; Ebizuka,Y. *Eur. J. Biochem.* **1998**, *256*, 238-244. (c) Kushiro, T.; Shibuya, M.; Ebizuka,Y. Molecular Cloning of Oxidosqualene Cyclase cDNA from Panax ginseng: The Isogene that Encodes β-Amyrin Synthase. International Symposium on Natural Medicines; **1998**, Kyoto Japan.

[51] Basyuni, M.; Oku, H.; Tsujimoto, E.; Kinjo, K.; Baba, S.; Takara, K. *FEBS J.* **2007**, *274*, 5028-5042.

[52] Hill, W.; Caballero, A.; Dempfle, L. *Genet. Sel. Evol.* **1996**, *28*, 379-383.

[53] (a) Crute, I. R. *Aspects Appl. Biol.* **1988**, *17*, 305-312. (b) Crute, I. R. *Euphytica* **1992**, *63*, 95-102.

[54] Niu, Y.; Luo, H.; Sun, C.; Yang, T. J.; Dong, L.; Huang, L.; Chen, S. *Gene* **2014**, *533*, 295-303.

[55] Xu, R.; Fazio, G. C.; Matsuda, S. P. T. *Phytochemistry* **2004**, *65*, 261-291.

[56] (a) Ebizuka, Y.; Katsube, Y.; Tsutsumi, T.; Kushiro, T.; Shibuya, M. *Pure Appl. Chem.* **2003**, *75*, 369-374. (b) Morlacchi, P.; Wilson, W. K.; Xiong, Q.; Bhaduri, A.; Sttivend, D.; Kolesnikova, M. D.; Matsuda, S. P. T. *Org. Lett.* **2009**, *11*, 2627-2630.

[57] Morita, M.; Shibuya, M.; Kushiro, T.; Masuda, K.; Ebizuka, Y. *Eur. J. Biochem.* **2000**, *267*, 3453-3460.

[58] Zhang, X.; Xiong, H.; Lui, L. *J. Ethnopharmacol.* **2012**, *141*, 206-211.

[59] (a) Kushiro, T.; Shibuya, M.; Ebizuka, Y. *Tetrahedron Lett.* **1999**, *40*, 5553-5556. (b) Shibuya, M.; Zhang, H.; Endo, A.; Shishikura, K.; Kushiro, T.; Ebizuka, Y. *Eur. J. Biochem.* **1999**, *206*, 302-307.

[60] Zhang, H.; Shibuya, M.; Yokota, S.; Ebizuka, Y. *Biol. Pharm. Bull.* **2003**, *26*, 642-650.

[61] Guhling, O.; Hobl, B.; Yeats, T.; Jetter, R. *Arch Biochem Biophys* **2006**, *448*, 60-72.

[62] Gauthier, C.; Legault, J.; Lebrun, M.; Dufour, P.; Pichette, A. *Bioorg. Med. Chem.* **2006**.

[63] Fotie, J.; Bohle, D. S.; Leimanis, M. L.; Georges, E.; Rukunga, G.; Nkengfack, A. E. *J. Nat. Prod.* **2006**, *69*, 62-67.

[64] Preetha, S. P.; Kanniappan, M.; Selvakumar, E.; Nagaraj, M.; Varalakshmi, P. *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **2006**, *143*, 333-339.

[65] Xue, Z.; Duan, L.; Liu, D.; Guo, J.; Ge, S.; Dicks, J.; OMaille. P.; Osbourn, A. *New Phytol.* **2012**, *193*, 1022-1038.

[66] Boutanaev, A. M.; Moses, T.; Zi, J.; Nelson, D. R.; Mugford, S. T.; Peters, R. J.; Osbourn, A. *Proc. Natl. Acad. Sci. USA* **2015**, 112, E81-E88.

[67] Inagaki, Y.; Etherington, G.; Geisler, K.; Field, B. Dokarry, M.; Ikeda, K.; Mutsukado, Y.;Dicks, J.; Osbourn, A. *New Phytol.* **2011**, *191*, 432-448.

[68] Pichersky, E.; Gang, D. R. *Trends Plant Sci.* **2000**, *5*, 439-445.

CHAPTER 6

Concluding Remarks

The metabolic diversity of natural products in higher plants appears to be driven out of their necessity to adapt and survive in different ecological environments. Among the vast amount of natural products found in nature, triterpenes are widely distributed across the plant kingdom. Triterpenes serve as precursors of essential membrane sterols in primary metabolism and have diverse roles in secondary metabolism, including defense and regulation.

The primary route used in this work for the study of triterpene biosynthesis and the discovery of novel OSCs uses genome mining strategies, where the availability of plant genomes and cDNA libraries allows identifying and isolating putative OSCs.

In this thesis, I have explored the differences and similarities in triterpene biosynthesis across species from the Brasicaceae and Astereaceae families using genome mining. A key component of this study was the knowledge that *Arabidopsis thaliana*, the model organism for the Brasicaceae, can provide in terms of OSC evolution, giving a foundation for detailed characterization of plant genes.[1]

Other powerful tools in the study of OSCs were the analytical methods developed in the course of the studying *A. thaliana*. Methods like GC-MS, ^{1}H NMR and 2D NMR gave this study a reliable process for quantification and characterization of complex mixtures.[2,3]

In Chapter 3 I described the role of the model organism *A. thaliana* in studying triterpene biosynthesis in the closely related species *A. lyrata*. The OSC analysis of *A. lyrata* was significantly easier than for *A. thaliana* because of the precedent established over the past decades in studying OSCs.[4]

Genomic DNA analysis, active site comparisons and protein expression were applied to determine the number of cyclases present and their possible product profile. At the end, nine OSCs were proposed to be functional *A. lyrata*. Due to the high level of orthology with *A. thaliana* OSCs, it was predicted that seven of them would likely make the same major products as their *A. thaliana* counterpart. This highly reduced the need for

protein expression. In the case where the cyclase was unique to *A. lyrata* (*Aly*PEN9), protein expression was required.

In Chapter 4, I applied the same core phylogenetic analysis and analytical techniques that were utilized throughout this thesis. However, the techniques used in studying *L. perennis* nematocyphol synthase (NEM) had to be adapted to suit EST mining strategies. With 40 products that ranged from 55% to <1%, NEM turned out to be one of the most product-diverse OSCs characterized by heterologous yeast expression. Nematocyphol, the major product, along with many minor NEM products, was found in the PEN6 product profiles of the *Arabidopsis* species described in Chapter 4. In addition, seven lupane type rearranged products were found for the first time to be part of the product profile of an OSC, including the novel compound lup-19(21)-en-3β-ol.

Finally, Chapter 5 presented an analysis of OSCs in the genome of *L. sativa* and product profiles of several *Lactuca* OSCs. Several genomes from the Asteraceae family have been studied, but only those from lettuce, horseweed, sunflower and globe artichoke are available, yet not completely annotated.[5] Thus the analysis of the genome in the *Lactuca* species was more challenging than in *A. lyrata*. The study of secondary metabolites by genome mining is limited in the Compositae family[6] due to the lack of available orthologs for the cyclases present in this family. This in turn complicated the prediction of the putative OSCs product profiles.

Gia Fazio[6] and the work presented in this thesis utilized EST mining to characterize six OSCs from *L.sativa*, *L. serriola* and *L. perennis*. The work done in these cyclases showed for the first time orthology between OSCs in the *Lactuca* family; moreover, it showed that OSC orthology is much higher in *Lactuca* than in *Arabidopsis*.

Genome mining in *Arabidopsis* and the *Lactuca* clade gave some insight into the similarities and differences in triterpene biosynthesis in distant organisms. The first variation highlighted by this study is the difference in orthology within genera. *A. thaliana* and *A. lyrata*, which diverged approximately 10 million years ago,[7] conserved their primary metabolism OSC (CAS) with high sequence similarity (98%) and identical active site. Orthology for the rest of the OSCs that have an ortholog is in the low 90s, thus making product profile prediction somehow less certain than in *Lactuca*. In addition, the number of OSCs differs between species. *A. thaliana* has 13 characterized OSCs,[4] while *A.*

lyrata only has 9 putatively functional OSCs, three of which (*Aly*PEN4, *Aly*PEN6 and *Aly*PEN9) I characterized by heterologous expression in yeast.

The genomes of these species also show significant differences. The genome of *A. thaliana* (125-Mb) is smaller than the one for *A. lyrata* (207-Mb); this downsizing of the large ancestral polyploidal genome mirrors a persistent selection for a smaller genome in Brasicaceae.[10] While the *Arabidopsis* genus displays significant sequence variation among OSC its orthologs, the *Lactuca* genus has a high level of orthology. Each pair of *Lactuca* orthologs shows a 99% sequence identity, as well as 100% active site sequence identity. This higher level of homology can be attributed to the younger age of *Lactuca* and to the high level of inbreeding[8] of cultivated *Lactuca* species in order to retain certain traits like taste and pathogen resistance from *L. serriola*.[9] Even with the high level of homology among *Lactuca* OSCs from different species, the analysis of OSCs and their product profile is still dependent on the availability of orthologous genes that were previously expressed (see Chapter 5).

Finally, the idea behind this work was to investigate whether the phylogenetic analysis of OSCs will be sufficient to predict the catalytic outcome of an OSC. To a certain extent, it is safe to say that if the cyclase has an ortholog and the active site is identical then it is possible to predict the major product and to expect a similar product profile. Even when an ortholog is not available, heterologous expression is still a powerful tool for studying novel OSCs. Comparison between triterpene biosynthesis in different organisms showed that *Arabidopsis* and *Lactuca* do share a high identity in terms of triterpene structures made by their OSCs. However, apart from primary metabolism (CAS), they do not have orthologs; thus, we still depend upon heterologous protein expression to predict product profiles of OSCs from species to species, even with the presence of model organisms within the genus.

REFERENCES

[1] Xue, Z.; Duan, L.; Liu, D.; Guo, J.; Ge, S.; Dicks, J.; OMaille. P.; Osbourn, A. *New Phytol* **2012**, *193*, 1022-1038.

[2] Shan, H.; Wilson, K. W.; Castillo, D. A.; Matsuda, S. P. T. *Org. Lett.* **2015**, *17*, 3986-3989.

[3] Castillo, D.A.; Kolesnikova, M.D.; Matsuda, S.P.T. *J. Am. Chem. Soc.* **2013**, *135*, 5885-5894.

[4] (a) Corey, E. J.; Matsuda, S. P. T.; Bartel, B. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 11628-11632. (b) Kolesnikova, M. D.; Xiong, Q.; Lodeiro, S.; Hua, L.; Matsuda, S. P. T. *Arch. Biochem. Biophys.* **2006**, *447*, 87-95. (c) Herrera, J. B. R.; Bartel, B.; Wilson, W. K.; Matsuda, S. P. T. *Phytochemistry* **1998**, *49*, 1905-1911. (d) Kushiro, T.; Shibuya, M.; Masuda, K.; Ebizuka, Y. *Tetrahedron Lett.* **2000**, *41*, 7705-7710. (e) Kolesnikova, M. D.; Wilson, W. K.; Lynch, D. A.; Obermeyer, A. C.; Matsuda, S. P. T. *Org. Lett.* **2007**, *9*, 5223-5226. (f) Shibuya, M.; Katsube, Y.; Otsuka, M.; Zhang, H.; Tansakul, P.; Xiang, T.; Ebizuka, Y. *Plant Physiol Biochem.* **2009**, *47*, 26-30. (g) Ebizuka, Y.; Katsube, Y.; Tsutsumi, T.; Kushiro, T.; Shibuya, M. *Pure Appl. Chem.* **2003**, *75*, 369-374. (h) Xiang, T.; Shibuya, M.; Katsube, Y.; Tsutsumi, T.; Otsuka, M.; Zhang, H.; Masuda, K.; Ebizuka, Y. *Org. Lett.* **2006**, *8*, 2835-2838. (i) Lodeiro, S.; Xiong, Q.; Wilson, W. K.; Kolesnikova, M. D.; Onak, C. S.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2007**, *129*, 11213-11222. (j) Morlacchi, P.; Wilson, W. K.; Xiong, Q.; Bhaduri, A.; Sttivend, D.; Kolesnikova, M. D.; Matsuda, S. P. T. *Org. Lett.* **2009**, *11*, 2627-2630. (k) Xiong, Q.; Wilson, W. K.; Matsuda, S. P. T. *Angew. Chem. Intl. Ed.* **2006**, *45*, 1285-1288. (l) Shibuya, M.; Xiang, T.; Katsube, Y.; Otsuka, M.; Zhang, H.; Ebizuka, Y. *J. Am. Chem. Soc.* **2007**, *129*, 1450-1455. (m) Fazio, G. C.; Xu, R.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2004**, *126*, 5678-5679.

[5] Scaglione, D.; Reyes-Chin-Wo, S.; Acquadro, A.; Froenicke, L.; Portis, E.; Beitel, C.; Tironel, M.; Mauro, R.; Monaco, A. L.; Mauromicale, G.; Faccioli, P.; Cattivelli, L.; Rieseberg, L.; Michelmore, R.; Lanteri1, S. *Sci. Rep.* **2016**, *6*, 19427.

[6] Fazio, G. C. *Ph.D. Dissertation*, Rice University **2006**.

[7] Hu, T. T.; Pattyn, P.; Bakker E. G.; Cao. J.; Cheng J. F.;Clark' R. M.; Fahlgren, N.; Fawcett, J. A.; Grimwood, J.; Gundlach,H.; Haberer, G.; Hollister, J. D.; Ossowski, S.;

Ottilar, R. P.; Salamov, A. A.; Schneeberger, K.; Spannagl, M.; Wang, X.; Yang, L.; Nasrallah, M. E.; Bergelson, J.; Carrington, J. C.; Gaut, B. S.; Schmitz, J.; Mayer, K. F. X.; Van de Peer, Y.; Grigoriev, I. V.; Nordborg, M.; Weigel, D.; Guo, Y. *Nat. Genet.* **2011**, *43*, 476-481.

[8] Hill, W.; Caballero, A.; Dempfle, L. *Genet. Sel. Evol.* **1996**, *28*, 379-383.

[9] (a) Crute, I. R. *Aspects Appl. Biol.* **1988**, *17*, 305-312. (b) Crute, I. R. *Euphytica* **1992**, *63*, 95-102.

10 Johnston, J. S.; Pepper, A. E.; Hall, A. E.; Chen, Z. J.; Hodnett, G.; Drabek, J.; Lopez, R.; Price, H. J. *Ann. Bot.* **2005**, *95*, 229-235.

Appendices

Appendix A

A1. Triterpene Structures and Atom Numbering

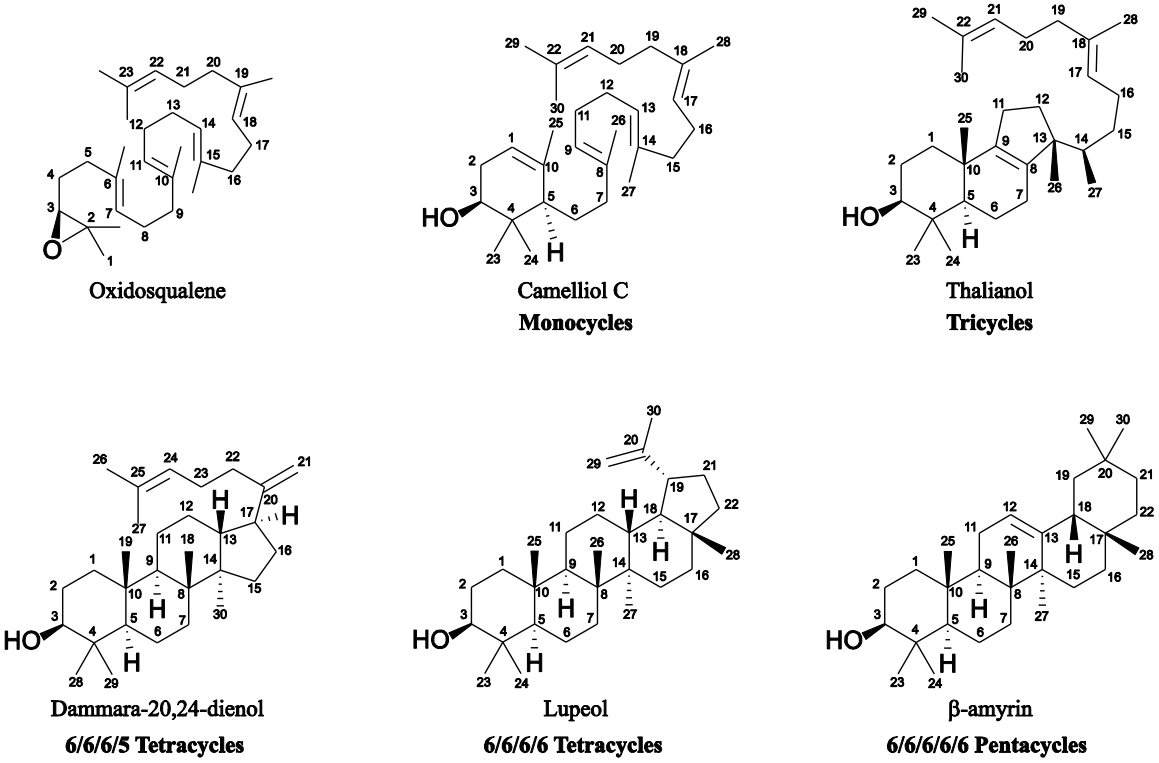


Figure A.1. Carbon atom numbering scheme used for triterpene structures. This atom numbering was used to describe the skeleton groups reported in this work.

A2. List of Abbreviations

*Aly*PEN4	thalianol synthase
*Aly*PEN6	seco-β-amyrin
*Aly*PEN9	β-amyrin synthase
Amp	ampicillin
BLAST	Basic Local Alignment Search Tool
BSTFA	N,O-bis(trimethylsilyl)trifluoroacetamide
CAS	cycloartenol synthase
$CDCl_3$	chloroform
cDNA	complementary deoxyribonucleic acid
COSYDEC	F1 decoupled 1H-1H correlation spectroscopy
DCM	dichloromethane, methylene chloride
DEPT	distortionless enhancement by polarization transfer
DEX	dextrose
DMAPP	dimethylallyl pyrophosphate
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphates
EDTA	ethylenediaminetetraacetic acid
EI	electron impact
ERG1	squalene epoxidase
ERG7	lanosterol synthase
EST	expressed sequence tag
EtOH	ethanol
FPP	farnesyl pyrophosphate
GC	gas chromatography, gas chromatograph
GC-MS	gas chromatography-mass spectrometry
GCP	Compositae Genome Project
GGPP	geranylgeranyl pyrophosphate
GPP	geranyl pyrophosphate
GTAE	Guanosine-Tris-acetate-EDTA

HMBC	heteronuclear multiple bond correlation
HPLC	high performance liquid chromatography
HSQC	heteronuclear single-quantum coherence
HSQC	heteronuclear single quantum coherence
Hsa ERG7	human lanosterol synthase
IPA	isopropanol
IPP	isopentenyl pyrophosphate
KOH	potassium hydroxide
LB	Luria-Bertani broth
LSS	lanosterol synthase
LUP1	lupane-3β,20-diol synthase
LUP2	b-amyrin synthase
LUP3	camelliol C synthase
LUP4	b-amyrin synthase
LUP5	tirucalla-7,24-dienol
MeOH	methanol
MS	mass spectroscopy, mass spectrometer
MSD	mass selective detector
NaOH	sodium hydroxide
NCBI	National Center for Biotechnology Information
NEM	*L. perennis* nematocyphol synthase
NMR	nuclear magnetic resonance spectroscopy
NOESY	Nuclear Overhauser effect spectroscopy
NSL	non-saponifiable lipid
OD_{600}	optical density at 600 nanometers
ORF	open reading frame
OS	2,3-oxidosqualene
OSC	oxidosqualene cyclase
OSCs	oxidosqualene cyclases
PCR	polymerase chain reaction
PEN1	arabidiol synthase

PEN2	baruol synthase
PEN3	tirucalla-7,24-dienol
PEN4	thalianol synthase
PEN5	marneral synthase
PEN6	seco-β-amyrin synthase
PMMA	polymethylmethacrylate
PTLC	preparative TLC
RNA	ribonucleic acid
RT	room temperature
SC	synthetic complete
SDS	sodium dodecyl sulfate
SOC	Super Optimal Broth with Catabolite Repression
SPE	solid phase extraction
SS DNA	single stranded salmon sperm deoxyribonucleic acid
TAE	40 mM tris base, 20 mM acetic acid, 1 mM EDTA
TE8	10 mM Tris-HCl, pH8.0, 0.1 mM EDTA
TIC	total ion chromatogram
TLC	thin-layer chromatography
T-MAS	testicular meiosis activating sterol, 4,4-dimethylzymosterol
TMS	tetramethylsilane (NMR), trimethylsilyl (GC-MS)
UV	ultra-violet
YP	yeast extract + peptone

Appendix B

B1. A*rabidopsis lyrata* growth conditions

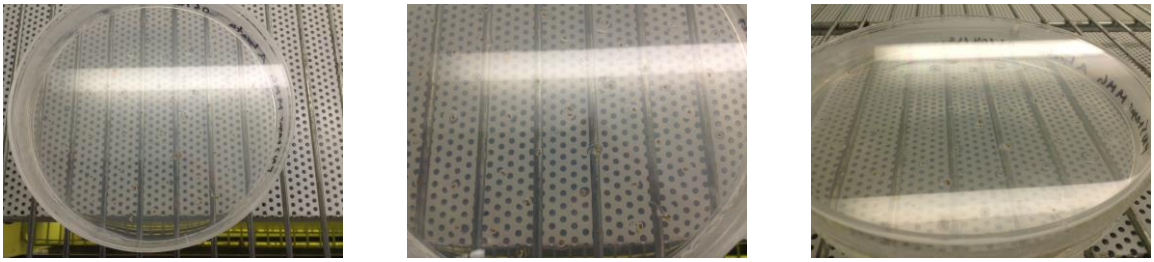


Figure B1.1 Seeds germinated after 7 days on PNS plates. After 20 days in PNS plates, seedlings were transplanted into soil following standard Arabidopsis growth procedures.

Figure B1.2. Two week's old rosettes from the time that seedlings were move to soil. Rosettes ready for vernalization at 4 ºC for four weeks.

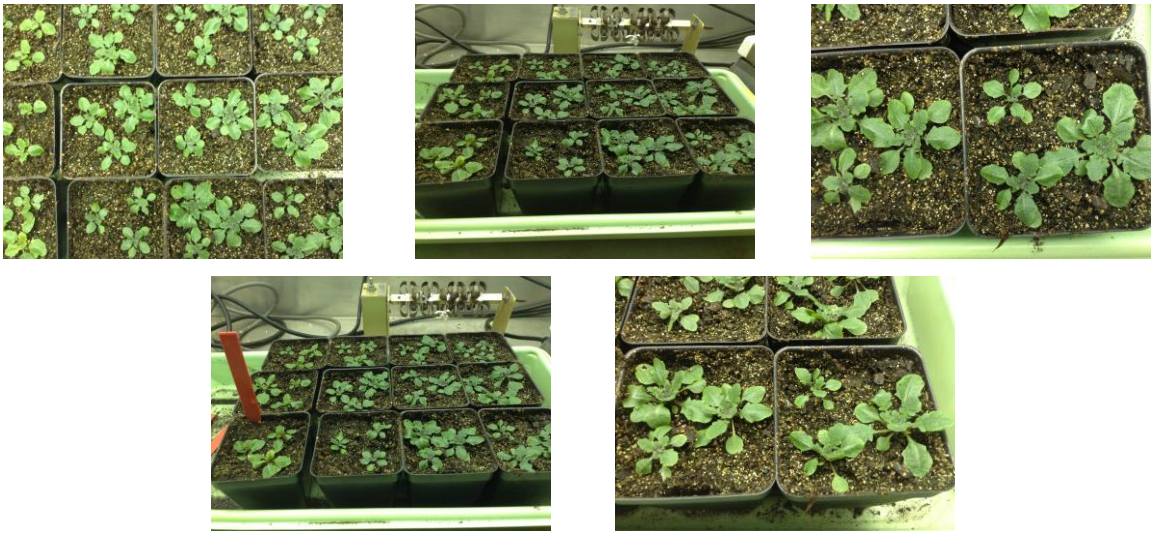


Figure B1.3. Fifteen days into verbalization of rosettes. The twenty day old seedlings were left in soil in the plant rooms for two weeks before the vernalization process started, so the plants could get use to the soil transplant.

Figure B1.4. Rosettes ready for bolting in the *Arabidopsis* plant rooms after four weeks in vernalization.

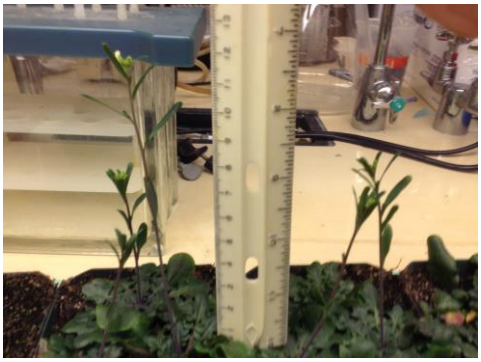


Figure B1.5. Plants starts to grow stem after 10 days in the *Arabidopsis* plant room. Three days after, the plants reach ~ 10 cm and start to produce flowers.

Figure B1.6. After two weeks from the time of bolting, flowers were collected for extraction. Flowers were collected daily for one week (cDNA extraction and plant extracts).

Figure B1.7. Four months old plants were harvested for RNA extraction and organic solvent extractions.

Appendix C

C1. Primers

C1.1 List of primers for RNA amplification of *Aly*PEN8.

PEN8cDNARTPCR	TTACGGACGCAGAGCTTGAG

C1.2 List of primers for sequencing of *DW017758 L. perennis* nematocyphol synthase.

T3	ATTAACCCTCACTAAAGGGA
T7	TAATACGACTCACTATAGGG
SP6	ATTTAGGTGACACTATAG
426GALF	AATATACCTCTATACTTTAACGTC
DW58Rint	GTAGACACATTAGTGCTTCAGCTGTAC

C1.3 List of primers for sequencing and PCR amplification of *L. perennis* nematocyphol synthase.

T3	ATTAACCCTCACTAAAGGGA
T7	TAATACGACTCACTATAGGG
426GALF	AATATACCTCTATACTTTAACGTC
pGCF18.5TCF/SalI	CCCGTCGACAAAAATGTGGGAGTTAAAG
pGCF18.5TCR	CCCGATATCGTATTCCACTCAAAATTGAAAGTTC
pGCF18.5NotI	CCCGATATCGTATTCCACTCAAAATTGAAAGTTC

C2. Sequences

C2.1 *Arabidopsis lyrata* putative OSCs.

Expressed sequences (*Aly*PEN4, *Aly*PEN6 and *Aly*PEN9) contain two restriction sites, a poly-A, and an additional stop codon. NCBI protein reference sequence name is provided, however online sequences are not proofreader; sequences provided in this document are.

C2.1.1 DNA sequence of XP_002883756 *A. lyrata* cycloartenol synthase (CAS).

ATGTGGAAACTGAAGATCGCGGAAGGAGGTAGTCCATGGCTTAGAACCACCAATAATCACGT
CGGAAGACAGTTTTGGGAGTTCGATCCGAATCTCGGTACTCCTGAGGATCTCGCCGCCGTCG
AGGAAGCTAGGAAGTCTTTTTCAGATAACCGATTCTTGCAGAAACATAGCTCCGATCTGCTT
ATGCGCCTTCAGTTTTCAAGAGAAAATTTGATTAGCCCAGTTTTACCTCAAGTCAAAATCGA
AGATACTGATGATGTTACAGAGGAGATGGTGGAAAACACGTTAAAGAGGGGTGTAGATTTCT
ATTCAACAATACAGGCGCACGACGGGCACTGGCCAGGTGATTATGGTGGTCCTATGTTTCTT
CTCCCAGGACTGATAATTACACTCTCCATAACTGGAGCACTGAATACAGTATTGTCGGAACA
ACATAAACAAGAAATGCGCCGTTATCTCTTTAATCACCAGAATGAGGACGGAGGTTGGGGTT
TACATATTGAGGGCCCTAGCACCATGTTTGGGTCTGTGTTGAACTATGTTACTCTAAGGTTG
CTTGGAGAAGGACCTAACGATGGAGATGGAGCTATGGAGAAAGGGCGAGACTGGATACTAAA
TCATGGTGGTGCTACCAATATCACATCTTGGGGGAAAATGTGGCTATCGGTACTTGGAGCTT
TTGAATGGTCTGGAAATAACCCACTGCCACCTGAGATATGGCTTCTTCCATATTTCCTGCCA
ATTCATCCAGGAAGGATGTGGTGCCATTGTCGGATGGTGTACTTGCCGATGTCGTATTTGTA
TGGAAAAAGGTTTGTGGGTCCCATAACGTCCACTGTTTTATCACTGAGAAAGGAGCTTTTCA
CTGTACCATATCATGAAGTCAACTGGAATGAAGCACGCAACCTTTGCGCAAAGGAGGATTTA
TACTACCCACACCCACTTGTGCAAGATATTCTTTGGGCATCACTTCATAAGATTGTTGAGCC
TGTTCTGATGCGATGGCCTGGTGCAAATTTGAGAGAAAAGGCTATAAGAACCGCAATAGAAC
ATATTCATTATGAAGATGAGAATACTAGGTACATCTGCATAGGTCCCGTGAACAAGGTATTA
AATATGCTTTGCTGTTGGGTAGAGGACCCAAACTCAGAGGCTTTCAAGTTGCACCTACCAAG
AATCCATGACTTTCTCTGGCTAGCTGAAGATGGAATGAAGATGCAGGGTTATAACGGAAGTC
AGCTATGGGATACAGGTTTTGCTATTCAAGCGATTTTGGCAACTAACCTTGTCGAAGAATAT
GGGCCCGTTTTGAAAAAAGCACATTCATTTGTCAAGAATTCTCAGGTGTTAGAAGACTGTCC
TGGAGATCTGAATTACTGGTACCGCCACATTTCTAAAGGGGCTTGGCCTTTCTCAACTGCAG
ATCACGGTTGGCCCATCTCCGACTGCACCGCAGAAGGACTGAAAGCTGCTCTTTTGTTATCC
AAAGTTCCCAAAGAGATTGTTGGTGAACCAATAGATGCAAAACGGTTATATGATGCTGTTAA
TGTTATCATTTCATTACAGAATGCAGATGGAGGCCTCGCAACATATGAGCTCACCAGGTCAT
ACCCTTGGTTGGAGCTAATCAACCCAGCAGAAACCTTTGGCGATATTGTTATAGATTATCCT
TACGTAGAATGTACATCAGCTGCTATCCAAGCTTTGATATCATTCCGAAAGCTGTATCCTGG
TCATCGGAAGAAGGAAGTAGATGAGTGCATTGAGAAGGCGGTTAAGTTCATTGAATCTATTC
AAGCAGCAGATGGCTCATGGTATGGATCATGGGCTGTTTGCTTCACATATGGTACGTGGTTT
GGAGTGAAAGGGCTGGTAGCTGTTGGAAAGACATTGAAAAACTCTTCACATGTTGCCAAAGC
TTGCGAATTTCTATTGTCTAAACAACAACCTTCGGGCGGATGGGGAGAAAGCTATCTTTCAT
GTCAAGACAAGGTATATTCAAACCTTGAAGGCAACAGGTCTCATGTCGTGAATACAGCATGG
GCTATGCTCGCACTCATTGGTGCTGGGCAAGCTGAGGTAGATCAAAAACCACTAGACCGGGC
TGCAAGATACTTGATTAATGCTCAGATGGAGAATGGTGATTTTCCACAACAGGAAATAATGG
GAGTCTTCAATAGGAACTGCATGATAACATATGCTGCGTATCGAAACATTTTTCCGATATGG
GCATTGGGGGAGTACCGTTGCCAGGTATTATTGCAACAAGGAGAATGA

C2.1.2 DNA sequence of XP_002877385 *A. lyrata* lanosterol synthase (LSS).

ATGTGGAGGTTAAAGTTATCGGAAGGAGACGAAGAGAGTGTGAATCAACATGTTGGAAGACA
GTTTTGGGAGTATGATAACCAATTTGGAACCTCTGAAGAGAGACATCACATCGCCGATCTTC
GGGGCAACTTTACTCTCAATCGGTTTTCTTCTAAGCATAGTTCTGATCTTCTCTACCGTTTT
CAGTGTTGGAAAGAGGAAGGAAAAGGAAAGGAGAGGCTTCCACAAGTGAAAGTAAAAGAGGG
GGGAGAAGGAGAGATAAATGAAGAAGTGGTGAATGTAACATTAAGAAGAAGTTTGAGATTCT
ACTCAACACTTCAATCACAAGATGGGTTTTGGCCTGGTGATTATGGTGGCCCTTTGTTTCTC
TCGCCTGCTCTGGTGATCAGCTTATATGTGACAGAAGTTTTAGACGCAACTTTAACTGCTCA
ACATCAAATGGAGATTCGTCGTTATCTCTATAACCATCAGAACAAGGATGGAGGATGGGGAC
TACACATAGAAGGGAGTAGCACCATGTTCTGTACGGCTCTCTCATACGTAGCACTAAGGCTC
ATGGGGGAAGAAATGGACGGTGGAGATGGAGCCATGGAATCAGCTAGACTTTGGATTCACCA
CCGTGGTGGTGCCACCTTTGTTCCCTCTTGGGGCAAGTTCTGGCTCTCCGTTCTTGGAGCTT
ATGAATGGAGTGGCAACAATCCTTTACCTCCAGAGCTATGGCTCCTTCCTTATAGTCTTCCA
TTTCATCCAGGCCGAATGTGGTGCCATTGTAGGATGGTTTATCTTCCAATGTCATATCTATA
CGGAAGAAGATTTGTTTGTCGCACTAATGGAACTATTTTGTCCCTTCGACGAGAGCTTTACA
CTGTTCCTTATCATCATATCGATTGGGACACGGCTCGTAATCAATGTGCCAAGGAGGACTTG
TACTATCCACATCCAAAGATTCAAGACGTTCTTTGGAGTTGTCTGAATAAATTTGGAGAGCC
TCTTCTTGAAAGATGGCCATTGAATAAGCTAAGAAGCCGGGCTCTTCAGACAGTGATGCAAC
ATATTCAATATGAAGACCAAAACAGCCACTATATTTGTATCGGTCCTGTCAATAAAGTCTTG
AATCTTCTTTGTTGTTGGGTTGATTCCTCGAATTCCGAGGCATTTAAATCTCACCTCTCGCG
TATTAAAGACTATTTATGGGTGGCTGAGGATGGAATGAAAATGCAGGGATACAACGGATCTC
AGCTGTGGGACGTGACTTTAGCGGTCCAAGCAATCTTGGCTACGAATTTGGTTGATGAGTAT
GATTTGATGCTTAAGAGAGCGCATAACTACATCAAGAACACTCAAATAAGGAAAGACACTTG
TGGAGATCCGGGGGTTGTGGTACCGACACCCGTGCAAGGGAGGATGGGGTTTCTCCACCGCAG
ACAATCCATGGCCTGTTTCTGACTGTACTGCTGAAGCCTTGAAGGCGTCGTTGTTATTGTCA
CAGATACCAGTTGATTTGGTTGGAGAAGCCATGCCTGAAGAGCATTTATTTGATGCTGTTGA
CTTTATCTTATCATTACAGAACAACAATGGGGGATTTGCGTCATACGAGCTAACTAGATCAT
ATCCCGCGCTAGAGGTGATCAATCCATCAGAGACTTTTGGGGATATCATCATAGATTATCAA
TACGTAGAATGTACGTCAGCTGCAATCCAAGGTCTTGCGTTATTCACAACGATCAATCCAAC
ATACAAGAGGAAAGAGATACTAATATGCATTAACAAAGCAGTTGAGTTTATTGAAAAAACAC
AACTTCCGGATGGTTCATGGTACGGCTCGTGGGGAGTGTGTTTCACCTATGCGACATGGTTT
GGTATTAAAGGGATGTTAGCTGCTGGCAAAACATATGAGACCAGTCTTTGTATTAGAAAAGC
TTGTGGTTTCTTGCTCTCCAAACAACTTTGTTGTGGTGGGTGGGGAGAGAGCTACCTTTCTT
GCCAAAACAAAGTATACACCAATCTTCCAGGGAACAAATCGCATATCGTGAACACATCATGG
GCAGTTTTGGCTCTCATTGAAGCAGGACAAGCTAATAGAGACCTGATGCCATTGCATCGCGG
GGCAAAATCGCTGATCAACTCGCAGATGGAAGACGGAGATTACCCTCAACAGGAGATATTGG
GAGTCTTTAATCGGAATTGTATGATCAGCTACTCAGCTTATAGAAACATATTCCCTATTTGG
GCTCTAGG

C2.1.3 DNA sequence of XP_ 002887764 *A. lyrata* lupeol synthase (LUP1).

ATGTGGAAGTTGAAGATAGGAAAGGGAAATGGTGAGGAACCGCATTTATTCAGTAGCAATAA
TTTCGTCGGACGTCAAACATGGGAGTTTGATCACCAAGCCGGCTCACCAGAGGAAAGAGCTT
CCGTCGAAGAAGCTCGCCGAAGTTTCTTGATCAACCGTTCTCGTGTTAAAGGTTGCAGTGAT
CTCTTATGGCGAATGCAATTTCTAAGAGAGAAGAAATTCGAACAAGGCATACCGAAACCTAC
TAAAATAAAGGAGGAAATAACGTATGAAACAACGACAAATGCATTACGAAGAGGCGTTCGTT
ACTTCTCAGCTTTGCAAGCCTCCGACGGCCATTGGCCTGGAGAAATCACTGGTCCGCTTTTC
TTCCTTCCTCCTCTCATATTTTGTTTGTACATTACCGGACATCTGGAGGAAGTATTCGATGC
TGAACATCGCAAAGAGATGCTACGGCATATCTATTGTCATCAGAACGAAGATGGTGGATGGG
GATCACATATCGAAAGCAAGAGTGTTATGTTCTGCACCGTGTTGAATTACATATGTTTTCGT
ATGCTTGGAGAAAATCCTGAACAAGACGCGTGCAAACGAGCTAGACAATGGATTCTTGACCG
TGGTGGTGTGATTTTTATTCCTTCTTGGGGGAAATTTTGGCTCTCGATACTCGGAGTTTATG

AATGGTCTGGAACTAATCCGACGCCACCAGAAATCTTGATGCTGCCTTCTTTTCTTCCAATA
CATCCTGGGAGAACTTTGTGTTATAGCCGGATGGTTAGTATACCTATGTCTTACCTATATGG
GAAGAGGTTTGTTGGTCCAATAACACCTCTTATTTTGCTCTTGCGCGGGGAACTTTACTTGG
AATCTTATGAAGAAATCAGTTGGAATAAAACTCGACGTCTATATGCAAAAGAAGACATGTAT
TATCCTCATCCTTTGGTTCAAGATTTGATATCTGACACTCTTCACAACTTTGTGGAGCCTTT
TCTTACACGTTGGCCATTGAACAAGCTTGTGAGGGAAAAAGCTCTTCAGCTGACTATGAAAC
ACATACATTATGAAGACGAAAATAGCCATTACATAACCATTGGATGTGTTGAGAAGGTACTG
TGCATGCTAGCTTGTTGGGTCGAAAATCCTAATGGAGACTACTTCAAGAAGCATCTGGCTAG
AATTCCAGATTACATGTGGGTCGCTGAAGATGGAATGAAAATGCAGAGCTTTGGATGTCAAC
TGTGGGATACCGGATTTGCTATTCAAGCTTTACTTGCAAGTAATCTCCCTGATGAAACTGAT
GATGCATTAAGGAGAGGACATAATTACATCAAGACATCTCAGGTTAGAGAAAACCCTTCAGG
TGATTTTAAGAGCATGTACCGCCACATTTCGAAAGGAGCATGGACATTTTCTGATCGAGATC
ATGGATGGCAAGTTTCAGATTGTACAGCTGAAGCTTTAAAGTGTTGCCTACTTCTTTCCATG
ATGCCAGCTGATATCGTTGGCCAGATAATAGATGATGAACAATTATATGATTCTGTTAATCT
CTTGCTATCTTTACAGAGCGGAAATGGAGGTGTCAATGCGTGGGAGCCTACCCGTGCATATG
AATGGATGGAACTGCTCAATCCTACAGAATTCATGGCTAATACCATGGTCGAGCGGAAGTTT
GTGGAATGCACCTCATCTGTTATACAAGCACTTGATCTATTTAGAAAGTTGTATCCAGATCA
CAGGACAAAAGAGATCAACAAGTCCATCAAAAAAGCTGTGCAATTTATACAAGGCAAACAAA
CAGCAGACGGTTCATGGTACGGAAATTGGGGTGTTTGCTTCATTTACGCTACTTGGTTTGCT
CTTGGAGGTCTAGCAGCAGCTGGTGAAACTTACAACGATTGTTTAGCTATGCGCAAAGGTGT
CCACTTTTTGCTCACTACACAAAGAGATGATGGAGGTTGGGGTGAAAGCTATTTATCATGCT
CCGAACAGAGATACATACCATTAGAAGGAGAAAGATCAAACATTGTGCAAACATCATGGGCT
ATGATGGCTCTAATTCATACGGGACAGGCTGAGAGAGATTTGATTCCTCTTCATCGTGCTGC
CAAACTAATCATCAATTCACAACTTGAAAACGGGGATTTTCCTCAACAGGAAATAGTAGGAG
CGTTCATGAATACATGCATGCTACACTATGCTACATACCGAAACACCTTCCCCTTATGGGCA
CTCGCGGAATACCGAAAAGTAATTTTAATTAATTAATCATCATTTACTTATTTATGTAATTT
T

C2.1.4 DNA sequence of XP_ 00288776 *A. lyrata* β-amyrin synthase (LUP2).

ATGTGGAAGTTGAAGATAGGAGAGGGAAATGGAGAAGACCCTTACTTATTCAGCAGCAACAA
CTTCGTCGGACGTCAAACATGGGAATTTGATCCCAAAGCCGGCACACCAGAGGAACGAGCCG
CCGTCGAAGAAGCTCGCCGGAAATTTTTAGACAACCGTTCCCGTGTTAAAGGTTGCAGTGAT
CTCTTGTGGCGAATGCAAGTGATTTTTGAAAGAGGCGAAATTCGAGCAAGTGATCCCGCCGG
TGAAGATCGACGACGGCGACGCCATAACTTACGAAAACGCGACGAATGCGCTACGGAGAGCA
GTTTCTTTCTTCTCGGCTTTGCCTCCGATGGCCACTGGCCGGCGGAAATCGCAGGAACTCTC
TTCTTCCTCCCTCCATTGGTTTTTTGTTTGTATATCACAGGACACCTCGAGAAGATATTCGA
TGCGGAACATCGCAGCGAGATGCTTCGGCATATCTATTGTCACCAGAACGAAGACGGTGGAT
GGGGATTACATGTTGAGGGAAACAGCGTTATGTTTTGCACCGCACTGAATTACATATGCTTG
CGTATGCTCGGAGAAGGTCCCAACGGAGGGCGAGATAACTCCTGCAAACGGGCCAGGCAATG
GATTCTTGACCATGGTGGTGTGACTTATATTCCTTCTTGGGGAAAATTTTGGCTCTCGATAC
TCGGAATCTATGATTGGTCTGGAACCAACCCAATGCCTCCCGAGATTTGGTTGCTGCCTTCT
TTCATTCCAATACACTTAGGAAAAACTTTGTGTTATACTCGAATGGTTTATATGCCCATGTC
TTATCTGTATGGGAAACGATTTGTTGGTCCTCTTACACCTCTTATCATGCTATTGCGCAAAG
AACTCCACTTACAACCTTATGAGGAAATCAATTGGAACAAAGCGCGCCGTCTATGTGCAAAA
GAAGACATGATTTATCCTCATCCTCTGGTTCAAGATTTGTTATGGGACACTCTTCACAATTT
TGTGGAGCCTTTCCTTACAAGTTGGCCGTTAAAAAAACTTGTACGGGAAAAGGCTCTTCGAG
TGGCAATGGAACACATACATTATGAGGACGAAAATAGCCATTATATTACCATCGGATGTGTT
GAGAAGGTTCTGTGCATGCTTGCTTGCTGGATCGAGAATCCTAATGGAGATCACTTTAAGAA
ACATCTCGCTAGAATTCCGGACTTCATGTGGGTTGCTGAAGACGGACTGAAAATGCAGAGCT
TTGGAAGTCAACTATGGGACACAGGGTTTGCGATTCAAGCTTTACTTGCTTGTGACCTTTCT
GACGAAACTGATGATGTACTCAGGAAAGGACACAATTTCATAAAAAATTCTCAGGTTAGAGA
AAACCCTTCAGGTGACTTTAAGAGCATGTATCGCCATATTTCCAAAGGAGCCTGGACTTTGT

CTGATCGAGATCATGGATGGCAAGTGTCAGATTGCACAGCTGAAGCTTTGAAGTGTTGCATG
CTGCTCTCCTTGATGCCAGCTGAGGTCGTTGGCCAGAGAATAGATCCTGAACAACTATACGA
TTCTGTTAATCTCTTGTTATCGTTGCAGGGTGAAAAAGGAGGTTTGACTGCATGGGAGCCTG
TCCGTGCACAAGAATGGTTGGAATTACTCAATCCCACAGATTTTTTTACTAGTGTTATGGCT
GAACGCGAGTATGTAGAATGTACCTCAGCTGTTATTCAAGCTTTGGTTCTATTCAAACAACT
TTATCCGGATCACAGAACGAAAGAGATCATCAAGTCGATTGAGAAAGGGGTGCAATTCATAG
AAAGCAAGCAAACGCCTGATGGTTCATGGTATGGAAATTGGGGTATCTGTTTTATCTACGCG
ACATGGTTTGCTCTGAGCGGGCTAGCAGCTGCTGGTAAAACTTACAAAAGTTGTCTGGCGAT
GCGCAAAGGTGTAGATTTCCTGCTTACGATACAAGAAGAAGATGGAGGTTGGGGCGAAAGCC
ATCTGTCATGCCCTGAGCAGAGATACATACCATTAGAAGGGAACAGATCAAACCTAGTGCAA
ACTGCTTGGGCTATGATGGGTTTGATTCATGCCGGACAGGCCGAGAGAGATCCTACACCTCT
TCATCGTGCTGCGAAACTTATCATCACTTCGCAACTTGAAAATGGGGACTTTCCGCAACAGG
AAATATTAGGAGTATTCATGAATACATGCATGCTACACTATGCTACATACAGAAACACCTTC
CCCTTATGGGCACTCGCGGAATATCGAAAAGCTGCCTTCTTAACTCATCAAGATCTATAG

C2.1.5 DNA sequence of XP_ 002889222 *A. lyrata* camelliol synthase (LUP3).

ATGTGGAAGTTGAAGATAGCAAATGGAAACAAAGAAGAGCCCTACTTGTTCAGCACCAACAA
CTTCCTCGGAAGACAGACATGGGAGTTTGATCCGGACGCCGGCATAGCAGAGGAACTAGCTG
CCGTCGAAGAAGCTCGTCGGAAATATTTTGATGATCGTTTTCAGGTTAAAGCTAGCAGCGAT
CTCATATGGCGTATGCAGTTTTTAAAAGAGAAGAAATTCGAGCAAGTTATACCTCCGGTAAA
AGTTGAAGATGCCAACAAGATCACGAGCGAAATAGCTACAAATGCGTTAAGGAGAGGGGTCA
ATTTCTTATCGGCATTGCAGGCCACTGATGGACACTGGCCTGCAGAAAATGCTGGTCCTTTA
TTCTTCCTTCCTCCATTGGTTTTCTGTCTATTTGTCACTGGACATCTCCATGAGATATTCAC
TCAAGAGCATCGTCGAGAGATCCTCAGATACATCTACTGTCACCAGAATGAAGATGGTGGCT
GGGGATTACACATAGAAGGAGACAGCACCATGTTCTGCACCACACTAAACTATATTTGCATG
CGCATACTTGGAGAATCTCCTTTTGGAGGACCAGGAAACGCGTGCAGACGGGCCAGGGATTG
GATTCTTGACCATGGGGGTGCAACATACATACCCTCTTGGGGCAAAACTTGGCTTTCTATAC
TTGGTGTCTTTGATTGGTCAGGAAGCAACCCCATGCCTCCAGAGTTTTGGATCCTACCTTCG
TTTCTTCCAATACATCCAGCGAAAATGTGGTGTTACTGCCGGTTGGTTTACATGCCAATGTC
TTATCTTTACGGGAAGAGATTTGTTGGTCCAATAAGTCCTCTTATTCTGCAACTACGCGAAG
AAATTTACTTGCAGCCTTATGCAAAAATAAACTGGAACAGAGCACGCCATCTATGTGCAAAG
GAAGACGCGTACTGTCCTCATCCACAAATTCAAGATGTTATATGGGACTGTCTTTACATCTT
CACTGAGCCGTTTCTTACATGTTGGCCATTTAATAAGCTGCTTAGGGAAAAAGCTCTTGGGG
TGGCAATGAAACACATACATTATGAAGACGAAAATAGCCGTTATATTACCATTGGATGTGTT
GAAAAGGCATTATGCATGCTTGCCTGTTGGGTTGAGGATCCTAACGGAAGTCATTTCAAGAA
GCATCTTTTGAGGATTTCTGATTACTTGTGGATTGCAGAAGATGGGATGAAAATGCAGAGCT
TTGGAAGTCAATTATGGGATTCAGGATTTGCCCTCCAAGCTTTAGTTGCAAGTGATCTCGCG
AACGAAATCCCAGATGTACTCAGGAGAGGATATGACTTTTTAAAAAATTCTCAGGTTAGGGA
GAACCCTTCGGGTGACTTTACGAACATGTTCCGTCACATCTCTAAAGGGTCGTGGACTTTCT
CTGATCGAGACCATGGATGGCAAGCTTCCGACTGCACAGCCGAAGGGTTTAAGTGTTGCCTT
TTGCTTTCGATGATGCCACCTGACATTGTTGGCCCGAAAATGGATCCCGAACAGTTATATGA
GGCTGTTACTATCTTACTGTCTCTACAGAGTAAAAATGGAGGTGTAACTGCTTGGGAGCCTG
CCCGTGGACAAGAATGGTTGGAATTGCTAAATCCTACTGAAGTTTTTGCTGACATTGTGGTT
GAGCACGAGTACAATGAGTGTACTTCATCAGCAATCCAAGCTTTGATTCTGTTCAAGCAACT
ATATCCGAATCACAGGACAGCAGAGATCAACACTTCCATCAAGAAAGCCGTGCAATATATAG
AGAGCATACAAATGCATGATGGTTCATGGTACGGAAGCTGGGGAGTTTGCTTCACATACAGT
ACATGGTTTGGTCTGGGAGGCCTCGCAGCTGCTGGAAAGACGTACAACAACTGTTTGGCTAT
GCGTAAAGGCGTTCATTTCCTTCTCACAACTCAAAAAGATAATGGAGGTTGGGGTGAAAGCT
ACTTGTCATGTCCTAAAAAGAGATACATTCCAAGTGAAGGGGATAGATCAAACTTGGTGCAA
ACCTCTTGGGCAATGATGGGTCTACTTCATGCTGGACAGGCAGAGCGAGATCCGGCTCCTCT
TCACCGTGCTGCGAAGCTCTTAATCAACTCTCAACTGGAGAATGGCGATTTTCCTCAGCAGG
AGATAACTGGAGCTTTCATGAAGAACTGCTTGTTACACTATGCAGCATACAGAAACATCTTC

CCCGTGTGGGCACTCGCAGAGTATAGGAGACGAGTTCCGTTGCCATATGAAAACCTTGAACA
GAGAGAAGAAGTAAGATCCATCATAACGTCTCAACAGTCACAGGTCATTGACCATGATGTAG
GATCATAA

C2.1.6 DNA sequence of XP_ 002868209 *A. lyrata* baruol synthase (PEN2).

ATGTGGAGACTGAGAATTGGTGCTAAGGCCAGAGACGATACTCATTTGTTCACCACCAATAA
CTACGTTGGGAGGCAGATTTGGGAGTTTGATGCCGACGTAGGCTCTCCCGAGGAACTTTCCG
AGGTCGAGGAGGCTCGTCGGAATTTCTCCAACAACAGGTCGCGTTTCAAGGCTAGTGCTGAT
CTTCTATGGCGGATGCAGTTTCTTAGGGAGAAGAAGTTCGAACAGAAGATTCCTGGAGTGAG
AGTAGAGGATGCAGAGAAAATAACGTATGAAGATGCAAAGACTGCATTAAGAAGAGGATTAC
TCTATTTCACGGCCTTGCAAGCTGATGATGGACATTGGCCTGCTGAAAATGCTGGTTCCATA
TTCTTCAATGCCCCTTTTGTTATATGTCTTTACATCACTGGACATCTTGAGAAAATCTTCAC
TCACGAGCATCGTATAGAGTTACTTCGTTACATGTACAACCATCAGAACGAAGATGGTGGGT
GGGGATTACACGTAGAAAGCCCCAGCAATATGTTCTGCACAGTCATTAACTACGTATGTTTG
CGGATCTTGGGAGTAGAAGCTGGTCATGATGACCAAGGAAGCGCTTGTGCAAGGGCTCGTAA
GTGGATCCTCGACCATGGTGGTGCTACCTACTCGCCCTTAATAGGAAAAACTTGGCTTTCGG
TTCTTGGAGTGTATGATTGGTCTGGCTGCAAACCTATACCCCCTGAGTTCTGGTTCCTTCCT
TCTTGTTTCCCTGTTAATGGAGGGACTCTCTGGATTTATTTACGGGATATTTTTATGGGGTT
GTCATACTTGTATGGTAAAAAGTTTGTTGCTACCTCAACACCTCTCATTCTCCAGCTCCGAG
AAGAAATTTATCCCGAGCCTTACACAAAAATCGATTGGAAGCAAGCACGAAACCGATGTGCA
AAGGAAGATCTCTACTATCCACAATCATTTTTACAAGATTTATTTTGGAAAGGTGTTCACAT
ATTCTCAGAAAATATCCTAAATCGATGGCCTTTCAACAATACCATAAGACAAAGAGCTCTTC
GAACTACTATGGATCTCGTTCACTATCATGATGAAGCCACCAGATACATTACAGGCGGATCT
GTGCCGAAGGCCTTTCATATGCTTGCATGTTGGGTAGAAGACCCAGATAGTGACTATTTTAA
GAAACATCTTGCTCGAGTCCCTGATTTCATATGGATTGGCGAGGATGGTCTTAAAATTCAGT
CTTTTGGAAGCCAGTTGTGGGATACAGCCTTGTCGCTACATGTCCTTATAGACGGTTTCGAT
GATGATGTTGATGACGAAATTAGGTTAACGCTCTTTAAAGGATACGAGTATTTGAAGGAATC
TCAAGTTACAGAGAACCCTCCTGGCGATCACATGAAAATGTTCCGTCACATGGCGAAAGGTG
GATGGACATTTTCGGACCAAGATCAAGGATGGCCAGTTTCAGATTGTACTGCTGAGAGTTTA
GAGTGCTGTCTATTCTTCGAGAGCATGCCATCAGAGTTTATTGGCAAAAAAATGGCTGTGGA
GAAACTCTATGAAGCCGTTGATTTCCTTCTCTATTTGCAGAGTGATAATGGAGGTATAACGG
CATGGCAACCAGCGGATGGGAAAACTTGGTTAGAGTGGCTTAGTCCAGTGGAGTTTATTGAA
GACGCGGTTGTCGAGCATGAGTATGTAGAATGTACGGGGTCAGCAATTGCAGCATTGGCTAA
GTTCAATAAACAGTTTCCAGGATACAAAAAGGAAGAGGTTGAACGGTTTATAACCAAGGGGG
TGAAGTACATTGAGGACTTTCAAATGGTGGATGGTTCATGGTACGGAAATTGGGGAGTGTGT
TTCATCTATGGAACCTTCTTTGCGGTAAGAGGTCTTGTGGCTGCAGGGAAGTGTTACAATAA
CTGTGAGGCAATTCGTAGAGCAGTTCGTTTCATTCTAGACACACAAAACCTGGAGGGTGGCT
GGGGAGAAAGCTATCTCTCTTGTCCAACCAAGAAATATACTCCTTTGATAGGAAACAAGACA
AATGTGGTGAATACAGGACAAGCACTTATGGTTCTAATTGCGAGTGATCAAATGGAGAGAGA
TCCTTTGCCGGTTCATCGCGCTGCCAAAGTGTTGATTAATTCACAGTTGGATAATGGTGATT
TCCCGCAACAGGAAATAATGGGAGTTTTCAAGATGAATGTGATGCTCCATTTTCCTACCTAT
AGGAACATGTTCACTCTTTGGGCTCTCACACATTACACCAAGGCTCTGCGAGGGCTTTGA

C2.1.7 DNA sequence of XP_ 002863840 *A. lyrata* thalianol synthase (PEN4).

GTCGACAAAA**ATG**TGGAGATTGAGATTGGGTCCTAAGGCAGGTGAAGATACACACTTGTTTA
CTACTAATAACTATGCTGGTAGACAAATATGGGAATTTGATGCCAATGCTGGTTCCCCAGAA
GAATTAGCCGAAGTTGAAGACGCTAGACAAAACTTCTCTAACAACAGATCAAGATTCAAAAC
TAGTGCTGATTTGTTATGGAGAATGCAATTCTTGAGAGAAAAGAAATTCGAACAAAAGATTC
CTAGAGTAATAGTCGAAGATGCAAGAAAAATTAAGTATGAAGACGCCAAAACTGCTTTAAGA
AGAGGTTTGTTATACTTTACAGCATTGCAAGCCGATGACGGTCATTGGCCAGCCGAAAATTC

AGGTCCTAACTTTTATGCTCCACCTTTCTTAATCTGTTTGTACATCACCGGTAATTTGGAAA
AGATTTTTACTCCAGAACATGTTAAGGAATTGTTGAGACACATCTATAACATGCAAAACGAA
GATGGTGGTTGGGGTTTACATGTAGAATCCCACAGTGTCATGTTCTGTACTGTTATCAATTA
CATCTGCTTGAGAATAGTTGGTGAAGAAGCTTCTCATGATGACCAGGGTAACGGTTGTGCAA
AAGCCCATAAGTGGATTATGGATCACGGTGGTGCAACCTATACTCCATTGATCGGTAAAGCC
TTGTTGTCTGTTTTGGGTGTATACGATTGGTCAGGTTGCAATCCTATACCACCTGAATTTTG
GTTGTTACCATCTTCATTCCCTGTAAACGGTGGTACATTGTGGATCTATTTGAGAGATACCT
TTATGGGTTTGTCTTATTTGTACGGTAAAAAGTTCGTCGCTGCACCAACACCTTTGATATTG
AAGTTGAGAGAAGAATTGTATCCAGAACCTTACGCCAAGATAAATTGGACACAAACCAGAAA
CAGATGTGCTAAGGAAGATTTGTACTACCCAAGATCCTTTTTACAAGACTTGTTCTGGAAGA
GTGTTCATATGTTCTCTGAATCAATCTTGGATAGATGGCCTTTGAATAAGTTGATCAGAGAA
AGAGCTTTGAGATCTACTATGTCATTGATCCATTATCACGATGAATCTACTAGATACATCAC
AGGTGGTTGTTTACCAAAGGCATTTCATATGTTGGCCTGCTGGATAGAAGATCCTAAGTCAG
ACTACTTCAAAAAGCACTTGGCAAGAGTTAGAGAATACATATGGATCGGTGAAGATGGTTTG
AAGATCCAATCCTTCGGTAGTCAATTGTGGGACACATCCTTGAGTTTGCATGTATTGTTGGA
TGGTATAGACGAACACGATGTTGATGAAATCAGAACTACATTGGTTAAGGGTTACGATTACT
TGAAAAAGTCCCAAATCACCGAAAATCCAAGAGGTGACCATTTCCAAATGTTCAGACACATA
ACAAAGGGTGGTTGGACCTTCTCAGATCAAGACCAAGGTTGGCCAGTTTCCGATTGTACAGC
TGAAAGTTTGGAATGTTGCTTGTTTTTCGAATCTTTGCCTTCAGAATTGATCGGTAAAAAGA
TGGATGTTGGTAAATTGTACGATGCTGTAGACTATTTGTTGTACTTGCAATCTGATAACGGT
GGTATTGCCGCTTGGCAACCAGTTGATGGTAAAGCATGGTTAGAATGGTTGTCTCCTGTTGA
ATTCTTGGAAGATGCAGTTGTAGAATATGAATACGTAGAATGCACTGGTTCAGCCATTGTCG
CTTTGGCAAAGTTTAATAAGCAATTCCCAGAATACAAAAAGGCTGAAGTTAAGCAATTCATT
ACCAAAGGTGCAAAGTACATCGAAGATATGCAAACTGTTGACGGTTCATGGTATGGTAATTG
GGGTGTATGTTTCATCTATGGTACATTTTTCGCTGTCAGAGGTTTAGTTGCAGCCGGTAAAA
CCTATGGTAACTGCGAAGCTATCAGAAAGGCAGTTAGATTTTTGTTGGATACTCAAAATTTG
GAAGGTGGTTGGGGTGAATCCTTCTTGAGTTGTCCAAACAAAATCTATACACCTTTGAAGGG
TAACTCTACTAACGTCGTTCAAACAGGTCAAGCTTTGATCGTTTTGATCATGGCAGATCAAA
TGGAAAGAGACCCATTGCCTGTACATAGAGCTGCAAAGGTCTTGATTAATTCTCAATTGGAT
AACGGTGACTTTCCACAACAAGAAATCATGGGTACCTTTATGAGAACTGTTATGTTGCATTT
CCCTACATATAGAAATAC**TGA**GCGGCCGC

C2.1.8 DNA sequence of XP_ 002887743 *A. lyrata* seco-β-amyrin synthase (PEN6).

GTCGACAAAA**ATG**TGGAGGCTGAAGATCGGAGCTAAGGGCGGAGATGAGACTCACTTGTTCA
CCACCAACAACTACACCGGAAGACAAACTTGGGAGTTTGATGCCGATGCCTGCTCCCCAGAG
GAACTCGCTGAGGTCAATGAAGCTCGGCAGAATTTCTTCATAAACCGGTCACGTTTCAAGAT
CAGTGCTGATCTCCTTTGGCGAATGCAGTTTCTAAGGGAGAAGAAGTTCGAGCAGAAGATTC
CGAGAGTGGAAATAGGAGATGCGGAAAAGATAACGTACAAAGACGCAAAGTCGGCACTGAGA
AGAGGGATACTATATTTCAAGGCATTGCAAGCTGAGGATGGACATTGGCCTGCTGAAAACTC
TGGTTCCTTGTTCTTCGAAGCTCCCTTTGTCATATGCTTGTACATTACTGGACATCTGGAGA
AAATCTTCAGTTTGGAACATCGCAAGGAACTATTGCGCTACATGTACAACCATCAGAACGAA
GATGGTGGGTGGGGACTACACGTGGAAGGCCAAAGTGCTATGTTCTGCACGGTCATCAACTA
CATTTGCCTACGGATTTTTGGAGTAGAAGCAGATCATGATGATATTAAAGGAAGTGGTTGTG
CAAGGGCTCGTAAGTGGATCCTTGATCACGGTGGTGCTACATATACGCCCTTGATCGGAAAA
GTTTGGCTATCGGTCCTTGGAGTGTATGATTGGTCCGGTTGCAAACCCATACCGCCTGAGGT
GTGGATGCTCCCTTCTTTTTCGCCCTTCAACGGAGGTACTCTATGGATTTATTTCCGGGAGA
TTTTCATGGGCGTGTCATACTTGTATGGTAAGAAATTTGTAGCTACACCAACACCTCTCATT
TTACAGCTTCGGGAAGAGCTTTATCCTCAACCTTATGACAAAATCCTATGGAGTCAAGCTCG
GAATATGTGCGCAAAGGAAGATCTCTACTATCCACAGTCATTTTTACAAGAAATGTTTTGGA
AAGGTGTTCATACATTGTCAGAGACTATCCTGAGTCGATGGCCTTTCAACAAGCTCATAAGG
CAAAAAGCTCTTCGAACCACAATGGAGCTCCTACATTATCACGATGAAGCGAGCCGATACTT
TACCGGTGGATGTGTGCCAAAGCCGTTTCATATGCTTGCTTGTTGGGTAGAAGACCCCGACG

GTGATTATTTTAAGAAACATCTTGCTCGAGTCCCCGATTATATATGGATTGGAGAGGATGGT
CTGAAAATCCAATCTTTTGGTAGCCAATTGTGGGATACAGCATTCTCACTTCAAGTCATGTT
AGCTTATCAAGATGTTGATGCCGATGATAATGATATTGGATCAACGTTAATGAAAGGATACA
ATTTCTTGAATAAATCTCAGCTTACACAAAATCCTCCTGGTGACCATAGGAAAATGTTTAAA
GACATTGTGAAAGGAGGATGGACATTTTCGGACCAAGACCAAGGATGGCCTGTTTCCGATTG
TACTGCTGAGAGTTTAGAGTGTTGCCTTGTCTTTGGAAGCATGCCATCGGAGTTGATTGGCG
AGAAAATGGATGTGGAGAGGCTCTATGATGCCGTTAACTTACTTCTCTATTTTCAGAGCAAA
AATGGAGGCATAGCAGTGTGGGAGGCAGCTCGTGGAAGAACCTGGCTAGAGTGGCTTAGTCC
GGTGGAGTTTATGGAAGACACAATCGTCGAGCATGAGTATTTAGAATGCACGGGGTCTGCGA
TAGTGGCATTGGCACGGTTCTTGAAAGAGTATCCGGAGCACAGAAAGGAAGAGGTTGAAAGG
TTTATCAAGAGTGCTGTCAAATACATAGAAAGCTTTCAAATGCCCGATGGTTCGTGGTATGG
AAACTGGGGTGTCTGTTTCATGTATGGAACCTTTTTCGCGGTAAGAGGTCTAGTGGCTGCAG
GCAAGACGTACCAGAACTGTGAGCCGATTCGTAGAGCGGTTCAGTTCCTTCTGGAGACACAA
AACGTTCAAGGCGGTTGGGGTGAGAGTTATCTATCTTGCCCCAGGAAGAGATATACACCTCT
AGAGGGAAACAGAACCAATGTGGTGAATACGGGACAAGCAATGATGGTTTTGATTATGTGTG
GTCAGATGGAGAGAGACCCTTTGCCTGTTCATCGCGCAGCTAAAGTGTTGATCAATTCCCAG
ATGGATAATGGCGATTTTCCGCAAGAGGAAATAATGGGAGTGTTCAAGATGAATGTGATGAT
CAATTATCCAAACTATAG**TGA**GCGGCCGC

C2.1.9 DNA sequence of *A. lyrata* β-amyrin synthase (PEN9).

GTCGACAAAA**ATG**TGGAGATTGAGAATAGCTGCAGAAGCTGAAGACAAGACTCATTTGTTTA
CTACAAACAACTACGCAGGTAGACAAATCTGGGAATTCGATGCCAACACATGTTCTCCAGAA
GAATTGGCTGAAGTCAATAAGGTTAGACAAAACTTCTCTTCAAACCCATCTAGATTCAAGCC
TTCAGCAGATTTGTTATGGAGAATGCAATTCTTGAGAGAAAAGAAATTCGAACAAAAGATTC
CTAGAGTTATTATCGAAGACGCCCAAAAGATTACCTATGAAGATGCCAAGACTGCTTTAAGA
AGAGGTATTTTGTACTTTGCCGCTTTACAAGCAGATGACGGTCATTGGCCAGCCGAAAACTC
CGGTAGTTTGTTTTTCGAAGCACCTTTCGTTATCTCCTTGTACATCACTGGTCATTTGGATA
AGATTTTCAGTTTGGAACACAGAAAGGAATTGTTGAGATATTTGTACAACCATCAAAACGAT
GACGGTGGTTGGGGTATACACGTAGAATCTCAATCAGTCATGTTCTGTACAGTTATTAATTA
CATCTGCTTAAGAATCTTCGGTATTGAACCAGATCATGACGGTCAAGAATCAGTTTGTGCAA
GAGCCAGAAAATGGATTTTGGATCACGGTGGTGCTACATATACCCCATTAATAGGTAAAATC
TGGTTGTCCGTATTAGGTGTCTACGATTGGAGTGGTTGCAAGCCTATTCCACCTGAATTTTG
GATGTTGCCAGGTTCCAGTCCTATAAATGCAGGTACTTTGTGGATCTATTTCAGAGATATTT
TCATGGCCTTGTCTTATTTGTACGGTAAAAAGTTCGTTGCTTCACCAACACCTTTGATAGCA
CAATTGAGAGAAGAATTGTACCCACAACCTTACAATAAGATTAATTGGTCCCAAGCCAGACA
ATTATGTGCTAAGGAAGACTTGTATTACCCACAAAGTTTCGTTCAAGATTTGTTCTGGAAGT
CCGTTTACATGTTCAGTGAAAACGTATTGAACCAATGGCCTTTTAATAAGTTGATCAGACAA
AGAGCATTGAGAAAGGCCATGGAATTAATCCATTATCACGATGAAGCAACCAGATACATTAC
TGGTGGTTGTGTACAAAAACCATTTCATATGTTGGCTTGCTGGGTCGAAGATCCTGACGGTG
ACTATTTCAAAAAGCACTTAGCAAGAGTTCCAGACTACATTTGGGTTGGTGAAGATGGTTTG
AAGATCCAATCTTTCGGTTCACAATTATGGGACACTTCCTTGAGTTTACAAGTTATGTTGGC
AGCCGTTGCTGATGTAGATGACGAAGTCATTAGATCTACATTGGTTAAGGGTTACAAGTTCT
TGGAAAAATCACAATTAACCCAAAATCCACCTGGTGACCATATGAAGATGTTCAGACACATA
ACAAAGGGTGGTTGGACCTTCTCCGACCAAGATCAAGGTTGGCCAGTATCTGATTGTACTGC
AGAATCATTGGAATGTTGCTTGATCTTCGAATCTATGTCTTCAGAAATCATAGGTGAAAAGA
TGGACATTGAAAGATTGTACGATGCTGTTAACTTCTTGTTGTACTTACAATCCAAAAATGGT
GGTATAAGTGCTTGGGAACCAGCATTGGGTAAAACTTGGTTGGAATGGTTATCTCCTGTTGA
ATTCATGGAAAACACCACTATCGAACATGAATACGTAGAATGCACAGGTTCAGCTATAATCG
CTTTAGCAAGATTCAAACAACAATTCCCAAGACACAGAACTGAAGAAGTCGAAAGATTCATT
ACAAAGGGTGTTAAGTACATCGAATCTTTCCAAATGCCTGATGGTTCATGGTATGGTAACTA
CGGTGTCTGTTTTATGTATGGTACCTTTTTCGCCGTCAGAGGTTTGGTTGCTGCAGGTAAAA
CTTACTGCAATTCTGAACCAATCCATAGAGCTGTTCAATTCTTGTTAGAAACCCAAAACATT

GAAGGTGGTTGGGGTGAATCCTATTTGAGTTGTCCAAATAAGAAATACACACCTTTGGAAGG
TAATAAGACTAACGTTGTAAACACATCACAAGCTTTGATGGGTTTAATTATGGGTGGTCAAA
TGAAAAGAGATCCATTACCTGTACATAGAGCCGCTAAGGTCTTGATAAACTCTCAATTAGAC
AATGGTGACTTTCCACAAGAAGAAATCAGAGGTGTATTCAAAATGAACGTCTTGTTGCATTA
CCCTACATACAGAAACATTAA**TGA**GCGGCCGC

C2.2 DNA sequence of *DW017758 L. perennis* nematocyphol synthase.

The following sequence contains the *L. perennis* nematocyphol synthase gene, two restriction sites, a poly-A, and an additional stop codon:

GTCGACAAAA**ATG**TGGGAGTTAAAGATAGCCGAAGAGGATGGTCCCTATTTGTATAGTACTA
ACAACTTTGTTGGTAGACAATTCTGGAAATTTAATCCTGATGCTGGAACTCCAAAAGAGAAA
GAAGAAATTGAACAGGTTCGGCAAAACTTCAAAGATAATCGGAAAAATGGTGGACAACATGC
TTGTGGCGATCTCTTGATGCGGATGCAGCTTATGAAGGAAAATCAAATTGATCTTATGGACA
TAGCTCCAATAAGACTAAATGATGATGAACAAGTCACTTTTGAATCTGTAACAACTGCAGTG
AAGAAAGCGGTCCGACTAAACCGTGCAATCCAAGCAAAAGATGGTCACTGGCCTGCTGAAAA
TGCTGGCCCCTTGTTCTTCACTCCTCCACTTTTAATTGCTTTATACATAAGTGGTACCATTA
ATACAATCTTGAATGAAGAACACCACAAAGAGATGATACGGTATTTCTACAATCATCAGAAT
GAAGACGGGGGATGGGGGTTTTTTATCGAAGGCAAAAGTACGATGATTGGATCTGCATTAGT
CTATGTAGCCCTGCGTATACTAGGAGAAGGAAAAAATGGAGGAGATGGTGCAATGGACCGAG
GCCGCAAGTGGATACTTGACCATGGAGGTGCAACCTCCATTCCCTCCTGGGGAAAGCTCTAT
CTTTCGGTGCTTGGAGTGTATGATTGGGAGGGCTGTAACCCATTGCCACCAGAATTCTGGAT
TTTGCCATCATCATTTCCTTTTCATCCCGCTGAGATGTGGTGCTATTGTCGGACAACCTACA
TGCCTATGTCATATTTATATGGGAAAAGATTCCAAGGACCCATCACACCTCTTGTTTCATCA
TTGCGAAAAGAAATCTACCCCACCCCTTTTGAGCATATTAACTGGAATAAACAAAGGAATAA
TTGTTGTAAGGAGGACTGGTATTATCCGCATTCATTTCTTCAAGATGTTTTGTGGCATACCC
TTCACTACATTACTGAGCCTGTCCTTAAATATTGGCCATTTTCCAAACTACGTGGGAGATCG
CTTGATAGAGTTGTCGAGCTAATGCGCTATGAATCACAAGAGACTAGATACATGACCATAGG
ATGCATTGAAAAAAGTCTACAAATGATGTGTTGGTGGGCAGAGAATCCAAAAGGTGATGAGT
TCAAATATCACTTAGCCAGAGTTCCAGACTATTTATGGATTGCAGAAGATGGAATGACAATG
CATAGCTTTGGCAGTCAAGTGTGGGATTGTGTTTTTGCAACTCAAGCAATTATTGCAAGTAA
CATGACTGAAGAATACGGTGATTGTCTCAAAAAGGCACACTTTTATTTAAGAGAATCTCAGA
TAAAAGAAAATCCTTCAGGAGATTTCACTCGAATGTATCGACATATAACTAAAGGAGCATGG
GCCTTCTCCGATCAAGATCATGGATGGGCTGTCTCTGATTGTACAGCTGAAGCACTAATGTG
TCTACTTTTACTATCAAACATGCCGAAAGAAATTGTTGGAGAGAAAGTTGACAATGCCCGAC
TATACGAGGCAGTGAATCTTCTTCTTTACCTACAAAGTCCTATAAGCGGAGGATTTGCTATT
TGGGAGCCACCGATTCCAAAACCATTTCTACAGCTTCTTAATCCTTCAGAGATGTTTGCAGA
TATAGTTGTTGAGAAAGAGCATCTGGAACCCTCAGCTTGCATTATTGTAGCTCTGGTAGAGT
TCAACCGTGTCCATCCAAGACACAGAAAGAAAGAAATAGAACTTTCAATTTCGAATGGGATA
CGATATCTGGAGGAAACGCAATGGCATGATGGTTCATGGTATGGTTACTGGGGAATATGCTT
CCTATATGGAACATTCTTTGCATTAAGGGCTTTAAGTGCTGCTGGAAAGACATATGACAATA
ACGAAGCAGTCTGTAAAGGTGTCAAATTCTTACTTTCCAAACAGAATGAAGAGGGGGGTTGG
GGAGAGAGTCGCCTATCTTGTCCTACTGAGGTGTATACACCGTTGGATGGAAACCGAACAAA
TTTGGTGCAGACATCATGGGCTATGCTTGGTCTTATGTCCTGTGGGCAGGTGGAGAGAGATG
TGACGCCCTTAGATAAAGCAGCAAAGCTGTTGATTAACGCTCAAATGGATAACGGAGATTTT
CCTCAACAGGAAATTACTGGAGCCTGGATGAGGAATTGCACTYTGCACTATGCACAATACAG
GAATATTTTCCCACTCTGGGCACTTGGGGAGTACCGAAAACGTGTTTGG**TGA**GCGGCCGC

C2.3 DNA sequences of expressed *Lactuca* OSCs.

The following sequences contain each the *Lactuca* gene, two restriction sites, a poly-A, and an additional stop codon.

C2.3.1 DNA sequence of *BU000828 L. serriola* tirucalla-8, 24-dienol synthase.

GTCGACAAAA**ATG**TGGAAGTTAAAGATAGCAGAAGGTAACGATCCTTACTTATTTAGCACTA
ACAACTTTGTTGGACGCCAAATTTGGGAGTTTGATCCCGATGCTGGAACTCCTGAAGAGCAC
CAAGAAGTCGAAAACGCTCGTCAACTTTTCCTTAACCGTCAAAAGGAAGGTTTTCAAGCATC
AGGCGATTTACTCATGCGGATACAGTTAATCAAGGAGAATGAAATTGATATATTAAGCATAC
CACCAGCAAGGTTAGGGGAAAATGAGGAAGCGAATCAAGAAGCGGTAACAACCACGATAAGA
AAAGCAGTCAGATTTAATCGTGCCATCCAAACAAAAGATGGTCATTGGGCTACTGAACATGG
TGGCCCTTTATTTTTCACTCCTCCACTTATTATTATCTTATACATTAGTGGAGCCATCGATA
CACATTTAACAAAGGAGCACAAGAAAGAGATGAAACGCTTTATCTACAATCATCAAAATGAA
GATGGAGGATGGGGATTTCACATTGAAGGACATAGCACTATGTTTGTGTCCGTGTTAAGCTA
CATATCCCTACGACTTCTAGGGGAAGAAAAAGATGACAAAAATGTTGCACTTACTCGAGCAA
GAAAATGGATACTTGACCATGGTGGTGCAACCTATGTACCATCTTGGGGAAAACTTTATCTT
TCGGTGCTTGGTGTATATGAATGGGAAGGATGCAACCCAATACCACCAGAATTTTGGATTTT
CCCCGAGTTTCTACCCTTTCATCCAGGAAAAATGTGGTGTTATTGTCGAACGGCCTATATGC
CCATGTCATACTTGTACGGAAGAAAATTCCATGGCCCAATCACTGATCTTGTTCTTCAACTG
AGGCAAGAGATTTATCTTACCCCATACGATGAGATAAATTGGAATAAACAACGCCATAAATG
TTGTAAGGAAGATCTCTATTACCCTCACACAATAGTCCAAGATTTGATATGGGATGGTCTTT
ATTACTTAAGTGAGCCACTCTTCAAGTACTGGCCTTTTACAAAACTAAGAGAAAAAGCTCTC
AAAAGAACAATTGAGTTAACTCGTTATAACGCTGAAGAAAGCAGGTACATTACCATGGCAAG
TATTGAAAAGGGTTTTCAAATGATGTGTTGGTGGGCGGAGAACCCAAACGGGAATGAATTCA
AGCACCACCTTGCTAGATTACCGGATTACTTGTGGCTAGCAGAAGATGGAATGAAGTCGCAA
ACATTTGGTAGTCAATTATGGTGTAGTGCATTTGCAACTCAAGCAATAATCGCAAGTAATAT
GCCTGAAGAATATGGGGATTCTCTTAAAAAGGCCCATTTCTTTATCAAAGAATCTCAGGTTA
AGCAAAACCCAAAAGGAGATTTCACTAAAATGTGTCGACAGTTTAGTAAAGGATCATGGACA
TTCACAGATCAAGATCATGGATGGCCTGTCTCGGATTGCACAGGAGAAGCATTGAAATGCCT
ACTTTTATTATCCCAAATGCCCGAAGAAATTTCTGGAGAAAATGTAGATAACCAGCGACTGT
ATGATGCTGTTAACTTCCTTCTTTATGTACAAAGCCCTACAACAGGAGGTTTTGCTGTTTGG
GAAAAACCAATCCCACATCCATATTTACAGACGTTGAATCCTTCAGAAATGTTCGCCGACAT
TGTTGTTGAAAGAGAGCATGTTGAGTGCACAACTTCAGTGATGCAAGCTCTCATAGAGTTCA
AACACTTCCACCCCGGGCATCGGGAAAAAGAAATAGAAAAGGCTGTGGCAAATGCAGTGCGT
TATCTAGAGGATATACAATGGGAAGATGGTTCATGGTATGGTTATTGGGGGATATGTTTCAT
ATGTGGCACATTCTTTTCGTTAGGAGGCTTAGAATCAGCTGGAAAAACATATAATGATTGTG
AAGCAGTTCGCAAAGGAGCCAAGTTTTTACTCTCGATACAAAATGAAGAAGGTGGTTGGGGA
GAAAGCTACAAATCTTGCCCTAGTGAAGTCTACACACCACTAGATGGAAATAGAACTAATAT
AGTTCAAACAGCATGGGCTATGCTAGGTCTCATGTCATCTGGACAGGCTGAAAGAGATCCAA
CACCATTACACAAAGCAGCAAAGATCTTGATCAATGCACAGATGGATGATGGAGATTTTCCA
CAACAGGAGATGACTGGAACTGCAATGAGGAACTGCATTCTACATTATCCGTTATATAGGAA
TATATTCCCGTTATTGGCACTTTCGAAATATCGCAATATATTTTGGGCTACATAA**TGA**GCGG
CCGC

C2.3.2 DNA sequence of *BQ869693 L. sativa* mixed amyrin synthase.

GTCGACAAAA**ATG**TGGAAGTTAAAGATAGCAGAAGGAAATGATCATTATTTGTATTCCACCA
ACAATTTTGTTGGCCGGCAAATTTGGGAATTTGACCCCGATGCTGGAACTCCGGTAGAGCGA
CAAGAAGTTGAAGATGCACGTCAGTATTTTAGAGACAATCGAAGGGAAGGTGTTCATCCATG
TGGCGATCTGCTTATGCGGATGCAGTTGATCAAAGAAAACGGAATTGATTTATTTAGCATAC
CACCAAGAAGATTGGGAGTGAATGAGGAAGTAAATTATGATGCAGTGACAACATCAGTTAAA
AAAGCACTCCGATTAAACCGTGCAATTCAAGCAAAAGATGGTCATTGGCCTGCAGAAAATGC
GGGCCCTATGTTTTTCACTCCTCCACTCCTTATTGCTATGTACATCAGTGGAGCCATAGATA
CGCATTTAACCAAAGAACACAAGGAAGAAATGATACGTTATATCTACAACCACCAAAATGAA
GATGGAGGGTGGGGATTTTATATAGAAGGACACAGCACCATGATTGGATCTGCTTTAAGCTA
TGTAGCCCTACGGTTACTAGGAGAAGGACCCGATGATGGGAACGGTGCAGTGAGCCGAGCAA
GGGAGTGGATACTCGAACATGGCGGTGCAATGTCGATTCCTTCTTGGGGCAAAACTTATCTA
TCGGTACTTGGGGTGTATGAATGGGATGGATGCAATCCACTTCCGCCAGAATTTTGGCTTTT
CCCAGAAACTTTACCATATCATCCAGCAAAAATGTGGTGCTATTGTCGGACAACTTATATGC
CCATGTCATACTTGTATGGGAGAAAATTCCATGGTCCAATCACTGATCTTGTTTTGCAACTT
CGACAAGAAATTCATCCGATCCCATATGATAACATAAATTGGAATAAACAACGCCACAACTG
TTGCAAGGAAGATCTCTACTACCCTCATTCAACACTTCAAGATTTGTTGTGGGATAGTCTTA
ACTACTTCAGCGAGCCACTTCTTAAGCATTGGCCTTTTAAGAAATTAAGAGAAAAAGGTCTC
AAAAGAGCAGTTGAATTAATGCGATATAGTGCTGAAGAGAGTAGATATATCACTATCGGATG
TGTTGAAAAGAGCTTGCAAATGATGTGTTGGTGGGCAGAGAATCCAAACGGGGATGAATTCA
AGCATCATCTTGCTAGGGTTCCGGATTACTTATGGCTAGCAGAAGATGGAATGAAGATGCAA
AGTTTCGGGAGCCAAGTATGGGATTGTACACTTGCAACTCAAGCAATAATCGCTAGTGATAT
GGTTGAAGAATATGGGGATTCCCTTAAAAAAGCCCATTTTTATATAAAAGAATCCCAAATAA
AACAAAACCCATCTGGAGATTTTAGTAAAATGTGTCGACAGTTTACTAAAGGAGCATGGACT
TTCTCTGACCAAGATCAAGGTTGGGTTGTCTCAGATTGCACAGCTGAAGCACTTAAGTGTCT
TTTATTACTATCCCAAATGCCAGAGGAAATTTCAGGAGAAAAGGCTGATAATGAAAGATTAT
ATGAGGCTGTTAATGTCCTTCTTTACTTACAAAGTCCTATAAGTGGAGGTTTTGCTATTTGG
GAGCCACCTGTCCCTCAACCATATTTACAGATGTTGAATCCTTCGGAGATTTTTGCAGACAT
TGTTGTTGAGAAAGAGCATGTTGAGTGCACATCATCAATTATTCAAGCCCTTTTAGCCTTCA
AAAGATTGCACCCAGGTCATAGGGAGAAAGAAATTGAAATTTCTGTGGCAAAAGCAGTTGGT
TTTTTGGAGGAAAAACAATGGCATGATGGTTCTTGGTATGGTTATTGGGGAATATGTTTCCT
ATATGGCACATTTTTTACAATAGGAGGCTTAATTTCAGCTGGAAAAACATATAACAATAGTG
AATCGGTTCGTAAAGCAGTAAATTTTTTCCTTTCAACACAAAATGAAGAGGGAGGATGGGGA
GAAAGCATACAGTCTTGCCCTAGTGAAGTATACACACCACTGGATGGAAGTCGAACAAATTT
AGTTCAAACATCATGGGCTATGCTTGGCCTTATGTTATGTGGACAGGCTGAAAGAGATCCAA
CACCCTTGCATAAAGCAGCAAAGATATTGATTAATGCACAAATGGATAATGGAGATTTTCCT
CAACAGGAGATTACTGGAGTCTACATGAAGAATTGCATGCTGCATTATGCAGAGTACAGGAA
CATTTTCCCACTTTGGGCACTTGGGGAATATCGCAAACGTGTTTGGGTCAATTAA**TGA**GCGG
CCGC

C2.3.3 DNA sequence of *BQ997802 L. serriola* mixed taraxasterol synthase.

GTCGACAAAA**ATG**TGGAAGCTCAAAATAGGTGAAAAGAATGGGAAATTCAACATCGGTGATG
GAAATGGTGATGAATATTTATATAGCACCAATAACTTTGTGGGGAGACAAACTTGGGAGTTC
GACCCTGATGCAGGCACGCAGGAAGAACGTGAACAAGTTGAAAAGTTTCGAGAACAATTTTT
GATTAATAAGAAGAAGCTTGACATCAGTTGTTGTGCAGATCTGCTCATGCGAAATCAGCTTA
TTAAGGAAAGCGGGATCGATCTTCTGAGCGAACCTCCGGTGAGACTTGGAGAGGAGGAGGAT
GTGAACTTTGAAGCGGTGACAACCGCAGTTAAAAAGGCAGTTCGATTAAACCGTGCAATCCA
AGCTTGGGATGGTCATTGGCCGGCTGAAAATGCGGGTCCGCTCTTCTTCACTCCTCCCCTGA
TAATTGCCCTGTACATAAGTGGTACATTGGATACAATCCTAACAGAAGAACACAAGAAAGAG
ATGATCCGTTATATGTACATCCATCAAAATGAAGATGGAGGATGGGGATTCTATATATCAGG
CAGAAGTACAATGATAGGGAGTGTCTTGAACTATGTGGGTCTAAGACTTCTTGGAGAAACTT

CATCACCCAATGACAATGGTGATGGTGCGCTTGCTAAAGGCCGTAAATGGATACTTGATCAC
GGTGGCGCTACTTCGATTCCTTCGTGGGGCAAGGTGTATCTCTCGGCGCTTGGAGTGTATGA
ATGGGCAGGCTGCAACCCTCTTCCTCCAGAATTCTGGCTTTTCCCTTCCTTTTTGCCTTATC
ATCCTGCAAAAATGTGGTGTTACTGCAGAACCACATACATGCCAATGTCATACTTGTATGGT
AGAAGCTTTCATGGGCCTATTACAGACCTTGTTATATCATTGAGAGACGAGATTCATGTCAT
TCCTTACCACCAGATTAATTGGAACAAACAACGACATAATTGTTGCAAGGAGGATCTGTACT
ACCCCCACACCTACATCCAAGATCTCTTATGGGATGGTCTTCATTATTTTAGCGAGCCACTT
ATCACAAAATGGCCCTTTAAGAAGTTAAGAGAGAAGGGTCTCCAAAGAGTTTTGGAGCTAAT
GCAATATAACGCTGAGGAAGGACGTTACATAACCATGGGTTGCGTGGAAAAGGCTTTACAGA
TGATGTGTTTCTTTGCCCTAGATCCAAATGGAATTGACTTCAAACGACACCTTGCTAGACTG
CCGGATTACTTATGGATGGCGGAGGATGGTATGAAGATGCAAAGTTTTGGTAGCCAGTTATG
GGATTGCACTCTTGTGACTCAAGCAATTCTCTCGAGTGATATGGTTGACGAGTATGGAGATT
CACTAAAAAAAGCTCACTTTTACTTGAAAGAATCACAGATAAAAGAGAACCCAAAAGGAGAT
TTCGCAAATATGTGTCGCCAATTTACAAAAGGGGCATGGACTTTCTCCGATCAAGATCAAGG
TTGGGTCGTCTCAGATTGCACTGCCGAAGCTTTGAAGTGTTTAATGGCGTTGTCGCAAATGC
CACAAGAAATTGTCGGAGAAAAGGCTGAAGTGGAGCAATTATATGATGCTGTGAATGTCCTT
CTTTACCTACAAAGTCCACAAAGTGGTGGTTTTGCAATTTGGGAGGCACCAGTCCCAAAACC
ATATTTAGAGAAATTAAACCCTTCAGAACTTTTTGCAGACATAGTGGTGGAAAGAGAGCATG
TTGAATGTACAGGCTCCATAATTCAGGCTCTACAAACGTTCAAAGATCTACATCCAGGGCAT
CGTGAGAAAGAAATCGAAGTTGCTATTGAAAAAGGCATACACTTTTTGGAAAACAGGCAACA
AGAAAATGGTTCATGGTATGGTTATTGGGGTATATGTTATCTCTATGGCACATATTTTGTGC
TTCAAGGATTAGTATCTTGTGGGAAAACATATGAAAATAGTGAAGCAGTTCGAAAAGCTGTC
AATTTTTTCCTCTCAACACAGAACTCAGAAGGTGGTTGGGGAGAGAACTTTGAGTCATGCCC
ACAAGAGAAATTTATACCTTTGGAGGGGAACCGAACAAATTTGGTGCAAACTTCATGGGCAT
TGCTCGGTCTTTTATGTGGTGGACAGGCTGAAAGGGATCAGACACCATTACACAAGGCAGCA
AAATTACTTATAAATGGGCAAATGGATGATGGAGATTTTCCTCAACAAGAAATAACGGGAGT
GTACATGAAGAATTGCATGTTACATTATGCAGAATATAGGAACACTTTTCCCCTATGGGCAC
TCGCTGAGTATCGAAAGCGTGTTTGGTTGCCAAAACAAGAAAACTAA**TGA**GCGGCCGC

C2.3.4 DNA sequence of *BU003435 L. serriola* lupeol synthase.

GTCGACAAAA**ATG**TGGAAGCTGAAAATAGCAGAAGGTGGTGATGATGAGTGGCTGACCACCA
CCAACAACCATGTTGGCCGCCAACACTGGGAGTTTGATCCGGATGCAGGAACCGAAGAGGAA
CGTACTGAGATCGAGAAGATGCGTCTTCACTTCAAACTTAATCGTTTTCAATTCAAACAAAG
TGCCGACTTGTTAATGCGTACTCAACTAAGAAAGGAGAATGCAATAGATGAAAATGAAATAC
CGAAAGCAATAAAATTGAATGAAACAGAAGAAGTATCAAATGACGCTGTTACAACTACACTC
CGAAGAGCCATTAGCTTCTACTCTACTATTCAAGCCCATGATGGCCATTGGCCTGCTGAGTC
TGCTGGGCCTTTGTTCTTCCTTCCTCCACTGGTGATAGCACTGTATGTGACTGGAGCCATGA
ATGATATTCTAACACCTGCACATCAGCTAGAAATCAAACGCTACATATACAATCATCAGAAT
GAAGATGGAGGCTGGGGATTACATATAGAGGGACACAGCACAATGTTTGGATCAGTACTTAG
TTACATTACATTGAGAATGCTTGGAGAAGAAGCCAACAGTGTTGCAGAGGACATGGCAGTGG
TTAAAGGCCGTAAATGGATCCTTGACCATGGTGGTGCAGTTGGGATTCCTTCGTGGGGAAAG
TTTTGGCTCACGATACTTGGAGTATATGAGTGGGGAGGCTGTAATCCCATGCCACCCGAATT
TTGGCTGCTCCCTAAGTTTTTCCCGGTTCATCCAGGCAAAATGTTGTGTTACTGTCGCTTAG
TGTACATGCCCATGTCGTACTTATATGGCAAAAGATTTGTGGGAAGAATAACAAAGTTGGTT
CACTCACTAAGGCAAGAGCTTTATACGAACCCTTATGATGAAATTAATTGGAATAAAGCACG
AAATACTTGTGCAAAGGAAGATCTGTATTATCCACACCCTTTTGTTCAAGATATGGTATGGG
CCACACTTCATAATGTTGTTGAACCTATCTTAACTCGTCGGCCATTTTCCACACTACGAGAA
AAGGCACAAAAAGTTGCAATGGAGCATGTTCACTATGAAGATAAGAGTAGTAGATATCTTTG
CATTGGATGTGTGGAAAAGGTATTATGCTTGATTGCAACATGGGTGGAAGATCCAAATAGTG
ATGCATATAAGCGTCATCTTGCAAGAATTCCTGATTACTTTTGGGTTGCTGAGGATGGGATG
AAGATGCAGAGTTTTGGATGTCAAATGTGGGATGCAGCCTTTGCTATTCAAGCCATTTTTTC
AAGTAATCTAGTAGAGGAGTACGGCCAGACTCTTAAAAAGGCACACGAGTTTGTAAAAGCAT

CACAGGTTCGTGATAACCCCCCTGGAGATTTTAGTAAAATGTATAGACACACTTCTAAAGGT
GCATGGACGTTTTCCATACAAGACCATGGTTGGCAAGTTTCTGATTGTACAGCAGAAGGCTT
AAAGGTTGCACTTTTGTACTCCCAAATGAGCCCAGAACTTGTAGGCGAAAAACTTGAAACCG
AGCGTCTCTACGATGCTGTCAATGTTATTCTTTCATTACAGAGTGAAAATGGTGGATTTCCA
GCTTGGGAACCACAAAGGGCATATGCTTGGTTGGAGAAATTCAATCCGACTGAATTTTTTGA
AGATGTGTTGATTGAACGAGAGTATGTGGAATGCACTTCATCTGCAGTCCAAGGTTTGACAC
TCTTCAAGAAGTTGCACCCTGGACATAGAACCAAGGAGATCGAATATTGTATTTCAAAAGCA
GTGAAGTACGTCGAAGACATACAAGAAAGTGATGGCTCATGGTATGGTTGTTGGGGAATTTG
CTACACCTATGGGACATGGTTTGCAGTAGATGCACTAGTAGCTTGTGGGAAGAACTATCATA
ACAGTCCCACCCTTCAAAAAGCTTGCAAATTTCTCTTATCCAAACAACTTCCAGATGGTGGA
TGGGGAGAGAGTTATCTTTCAAGCTCAAACAAGGTGTATACGAATTTAGAGGGAAACCGGTC
GAATTTAGTACATACATCTTGGGCTTTGATATCACTTATTAAAGCTGGACAGGCTGAAATTG
ATCCTACACCAATAACTAATGGAATACGACTTCTCATCAATTCACAAATGGAAGAAGGAGAC
TTTCCTCAACAGGAAGTCACAGGAGTATTCATGAAGAACTGTAACCTCAATTATTCATCATA
TAGAAATATTTTCCCAATATGGGCACTTGGTGAGTATGGGCGTATTGTTCAAAATGTATGA**T**
GAGCGGCCGC

C2.3.5 DNA sequence of *BU003176 L. serriola* nematocyphol synthase.

GTCGACAAAA**ATG**TGGGAGTTAAAGATAGCCGAAGGGGATGGACCCTATTTGTTTAGTACTA
ACAACTTTGTTGGTAGACAATTTTGGGAATTTAATCCTAATGTTTGCACAACAAAAGAGAAA
GAAGAGATTGAAAAGATTCGACAAAATTTCAAAGATAATCGAAAAAATGGAGGACAACATGC
TTGTGGCGATCTCCTGATGCGGATGCAGCTTATGAAGGAAAATCAAATTGACCTTATGGACA
TAGCTCCAATTAGACTAACCGATGATGAACAAGTTACTTTTGAAGCCGCAACAACTGCAGTG
AAGAAAGCAGTCCGACTAAACCGTGCAATCCAAGCAAAAGATGGTCACTGGCCTGCTGAAAG
TGCTGGCCCCTTGTTCTTCACTCCTCCACTTTTAATTGCTTTGTACATAAGTGGTACTATTA
ATACAATCTTGAATGAAGAACACCACAAAGAGATGATACGGTATTTCTACAATCATCAGAAT
GAAGATGGAGGATGGGGGTTTTTTATCGAAGGCAAAAGCACGATGATTGGATCTACATTGGT
CTATGTAGCCTTGCGTATGCTAGGAGAAGGAAAAAATGGAGGAGATGGTGCAATGGACAGAG
GCCGCAAGTGGATACTTGACCATGGTGGTGCAACCGCCATTCCCTCCTGGGGAAAGCTCTAT
CTATCGGTGTTGGGAGTGTATGATTGGGAGGGTTGCAACCCATTGCCACCAGAATTCTGGAT
TATCCCATCGGCATTTCCTTTTCATCCTGCTGAGATGTGGTGTTATTGTCGGACAACCTACA
TGCCTATGTCATATTTATATGGGAAAAGAGTCCAAGGACCCATCACACCTCTTGTTTCATCA
TTGCGAAAAGAAATCTACCCCACTGCTTTTGAGCATATAAATTGGAATAAGCAACGAAATAA
TTGTTGTAAGGAGGACTTGTACTATCCACATTCATTTCTTCAAGATGTTTTGTGGCATACCC
TTCACTACATTACTGAACCTATCCTTAAATATTGGCCATTTTCCAAACTACGAGGGAAATCG
CTTGATAGAGTTGTTGAGCTAATGCGCTATGAATCAGAAGAGACTAGATACATGACCATAGG
ATGCATCGAAAAAAGTCTACAAATGATGTGTTGGTGGGCAGAGAATCCAAATGGCGATGAGT
TTAAATATCACTTAGCAAGAGTTCCAGATTACTTATGGGTTGCAGAAGATGGAATGACAATG
CATAGTTTTGGTAGTCAAGTGTGGGATTGTGTTTTTACAACTCCAGCAATTATTGCAAGTAA
CATGACTGAAGAATATGCTGATTGTCTCAATAAGGCACACTTTTATTTAAGAGAATCTCAGA
TAAAACAAAATCCTTCAGGAGATTTCACTCGAATGTATCGACATATCACTAAAGGAGCATGG
GCCTTCTCTGATCAAGATCATGGATGGGCTGTCTCTGACTGTACAGCTGAAGCACTAATGTG
CCTACTCTTACTATCAAACATGCCGAAAGAAATTGTTGGAGAGAAAGTTGACAATGCCCGAC
TATATGAGGCAGTGAATTTTCTTCTTTACTTGCAAAGTCCGATAAGCGGAGGATTTGCTGTT
TGGGAGCCACCGATTCCAAAACCATTTCTACAGCTTCTTAATCCTTCAGAGATGTTTGCAGA
TATAGTTGTTGAGAAAGAGCATTTGGAACCCTCAGCTTGCATTATTGTAGCTTTGGTAGAGT
ACAACCGCGTCCATCCAAAACACAGAAAGAAAGAAATAGAACTTtcAATTTTGAGTGGAATA
CGATATCTAGAGGAAACACAATGGCATGATGGTTCATGGTATGGTTACTGGGGAATATGCTT
CTTATATGGAACATTCTTTGCCTTAAGGGCTTTAAGTGCTGCAGGAAAGACATACGACAATA
ATGAAGCAGTATGTAAAGGTGTCAAGTTCTTACTTTCCAAACAGAATGAAGAGGGGGGTTGG
GGAGAGAGCCACCTATCTTGTTCTACTATGGTGTATACGCCATTGGATGGAAACCGAACAAA
TTTGGTGCAAACATCATGGGCTATGCTTGGTCTTATGTTCTGTGGGCAGGTGGAGAGAGATA

TAACGCCCTTAGATAAAGCAGCAAAACTGTTGATTAATGCTCAAATGGATAACGGAGATTTT
CCTCAACAGGAAATTACTGGAGCTTGGATGAGGAATTGCACTCTGCATTTTGCACAATACAG
GAGCATTTTCCCACTCTGGGCACTTGGGGAGTACCGTAAACGTGTTTGGTGA**TGA**GCGGCCG
C

C2.3.6 DNA sequence of *BU008606 L. serriola* cycloartenol synthase.

GTCGACAAAA**ATG**TGGAAACTGAAGATCGCAGAGGGAGGGAGTCCATGGCTGCGTTCGACAA
ACGATCACGTCGGCCGACAATTTTGGGAGTTTGATCCCACGCTAGGGTCCCTTGAGGAACTC
GCTGATATCGAGAAAGTTCGACAGACCTTTCATGAAAATCGGTTTGAGAAAAAACACAGTTC
AGATCTGCTTATGCGCAGTCAGTTTGCAAAAGAGAAGTCACTCTCTGTATTCCCACCTAGAG
TGAACATAAAAGATGCTGAAGATATCACAGAGGAGAAAGTAGCAAACGTCTTACGAAGAGCT
ATTGGTTTCCATTCAACCCTCCAGGCAGATGATGGGCATTGGCCAGGAGACTATGGAGGCCC
TATGTTTTTATTGCCTGGTTTGGTTATTACTCTAACTATTACTGGGGCACTGAATGCAGTCT
TATCTAAAGAGCATAAACGGGAGATGTGCCGTTACCTTTACAATCATCAGAATAGAGATGGT
GGATGGGGTCTACACATCGAGGGTCATAGCACCATGTTTGGTAGTGCCTTGAACTATGTCAC
TTTGAGATTGCTGGGTGAAGGAGCTAATGATGGAGAAGGGGGGGCGATGGAGAAGGGGCGGA
AGTGGATTTTGGATCATGGTGGTGCCACTTCAATAACATCTTGGGGAAAGTTTTGGCTTTCA
GTACTTGGGGTATTTGAGTGGTCTGGAAATAACCCGTTGCCCCCTGAGATGTGGGTCCTTCC
ATATTTCCTTCCCGTGCATCCAGGTAGGATGTGGTGTCACTGCCGGATGGTGTATCTGCCTA
TGTCATACTTATATGGGAAGAGGTTTGTTGGACCTATAACATCTACGGTTCTGGCCTTGAGA
AAGGAGCTGTTTACAGTCCCCTATCACGACATAGATTGGAATGATGCACGCAACCTCTGTGC
CAAGGAAGATCTTTACTACCCTCACCCACTCATCCAAGACATACTTTGGGCAACTCTTGACA
AGTTTGTGGAGCCTATACTGATGAGATGGCCTGGGAAGAAGCTGAGAGAGAAGGCTCTTCGC
ACTGCAATGGAACACATCCATTATGAGGATGATAATACTCGCTATATATGCATAGGGCCTGT
AAACAAGGTGTTAAATATGCTCTGCTGCTGGGCAGAAGATCCAAACTCAGAGGCTTTCAAGC
TACACCTCCCAAGGATACATGATTATCTTTGGCTTGCTGAAGATGGCATGAAAATGCAGGGT
TACAATGGTAGTCAACTTTGGGATACCGCATTCACCGTTCAAGCAATTATTTCTACAAACCT
CATCGAAGAGTTTGGTCCAACACTGAAAAAAGGACACATGTTCTTAAAGAAGTCACAGGTGT
TGGATAACTGCCCTGGcgATCTTGATTATTGGTATCGTCATATTTCAAAAGGTGCTTGGCCT
TTCTCAACAGCAGATCACGGATGGCCCATTTCAGATTGTACTGCAGAAGGGCTCAAGGCGGC
ACTCTTGCTTTCAAAATTGCCATCAAAAATCGTTGACGAGCCATTGGATGCAAAACGGCTGT
ATGATGCTGTCAATGTTATTTTATCTTTACAGAATTCTGATGGTAGTTTTGCAACATATGAA
CTTACAAGATCCTACAGCTGGTTGGAGTTGGTGAATCCTGCTGAAACCTTTGGTGACATTGT
TATTGACTACCCATATGTAGAGTGCACCTCGGCTGCGATTCAAGCTCTGGTGGCATTCAAGA
GATTGTACCCTGGGCATAGGCGAGAAGAGGTACAGCGTTGTATTGATAAATCTGCCTCCTTT
ATCGAAAAAATCCAAGCACCAGATGGTTCATGGTATGGTTCGTGGGCAGTTTGTTTCACGTA
TGGCACATGGTTTGGAGTGAAAGGCTTAGTGGCTGCTGGAAGGAACTTCTCTAACTGCTTTA
GCATTCGCAAGGCTTGTAACTTTTTGTTGTCCAAACAACTTGCTTCTGGGGGATGGGGGGAG
AGCTACCTCTCTTGTCAGAACAAGGTGTACACCAATCTTGAGGGAAATCGATCTCATGTGGT
AAACACAGGATGGGCTATGCTAGCTCTCATTGATGCTGAGCAGGCCAAGAGAGATCCAACAC
CATTGCATCGTGCAGCAAGAGTATTGATTAATTCTCAGATGGAAAATGGAGATTTTCCACAA
CAGGAGATCATGGGGGTTTTTAACAGGAATTGCATGATTACTTATGCTGCCTACAGAAACAT
CTTCCCTATTTGGGCATTAGGAGAATACAGATGTCGAGTACTTGAGGAGAAATCCATGCTAA
GCTAA**TGA**GCGGCCGC

Appendix D

D1. Additional Spectral Data

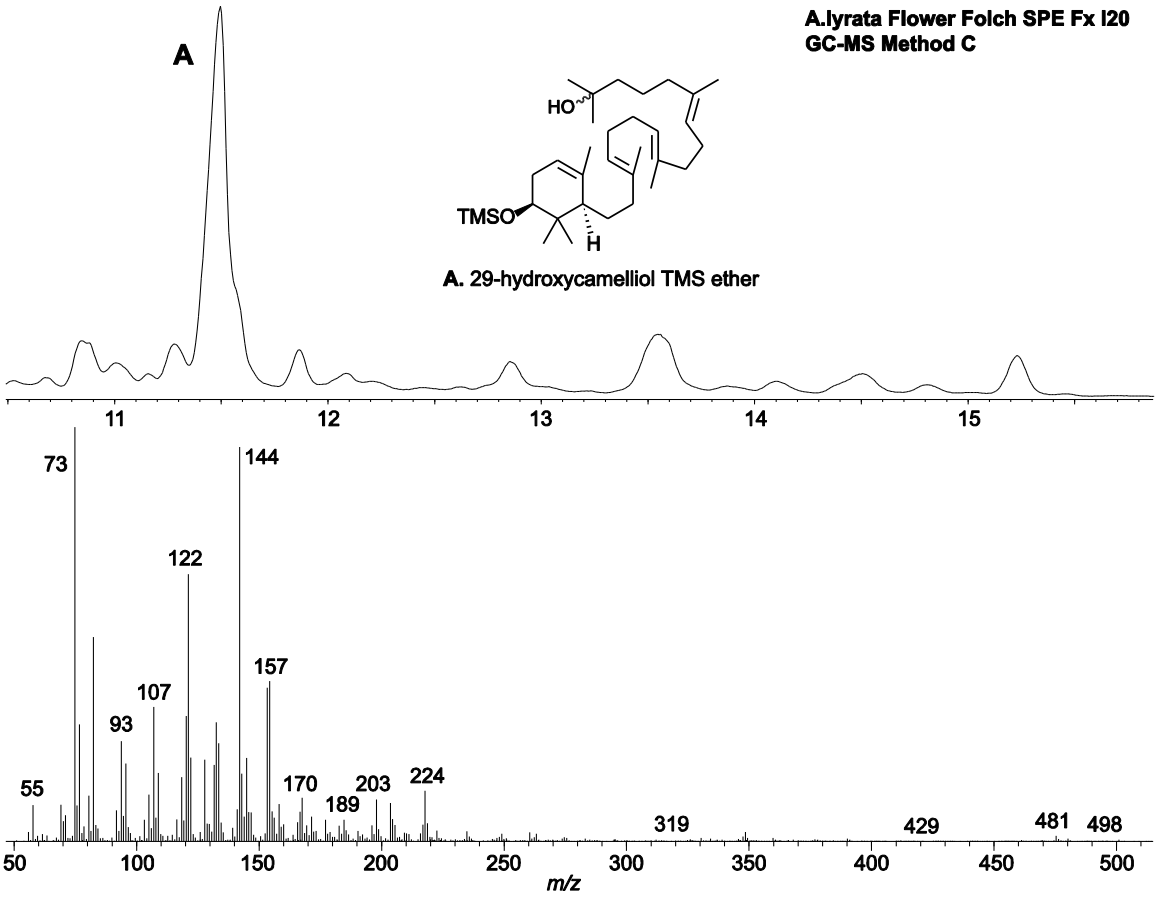


Figure D1.1. Expanded portion of the GC-MS chromatogram of *A.lyrata* Flower-Folch Extraction-SPE Fraction I20, indicating the presence of 29-hydroxycamelliol.

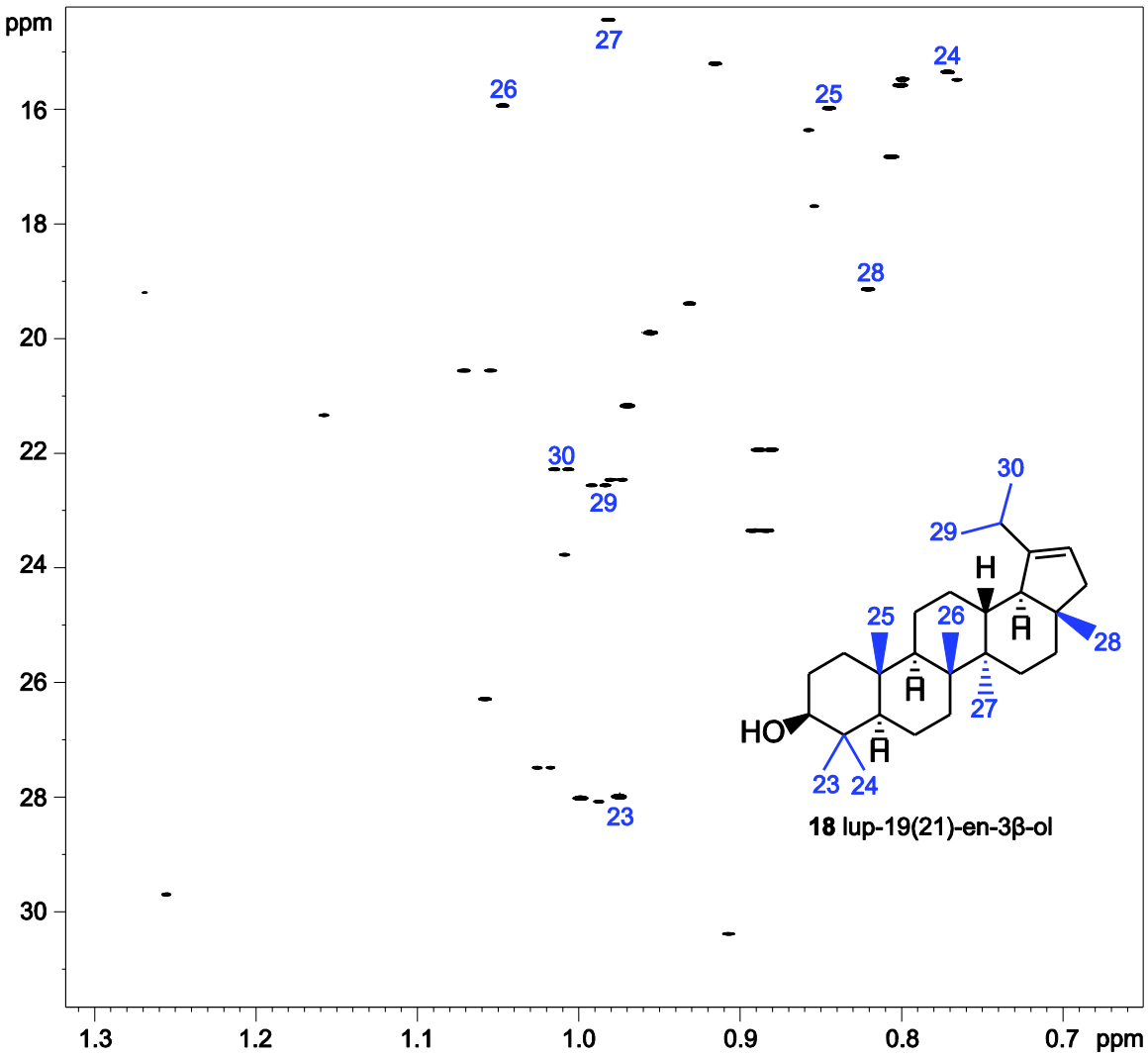


Figure D1.2. HSQC spectra of lup-19(21)-en-3β-ol.

Appendix E

E1. HSQC NMR Spectral Data

E1.1 HSQC spectra of *in vivo Aly*PEN4 SPE Fractions 22-35.

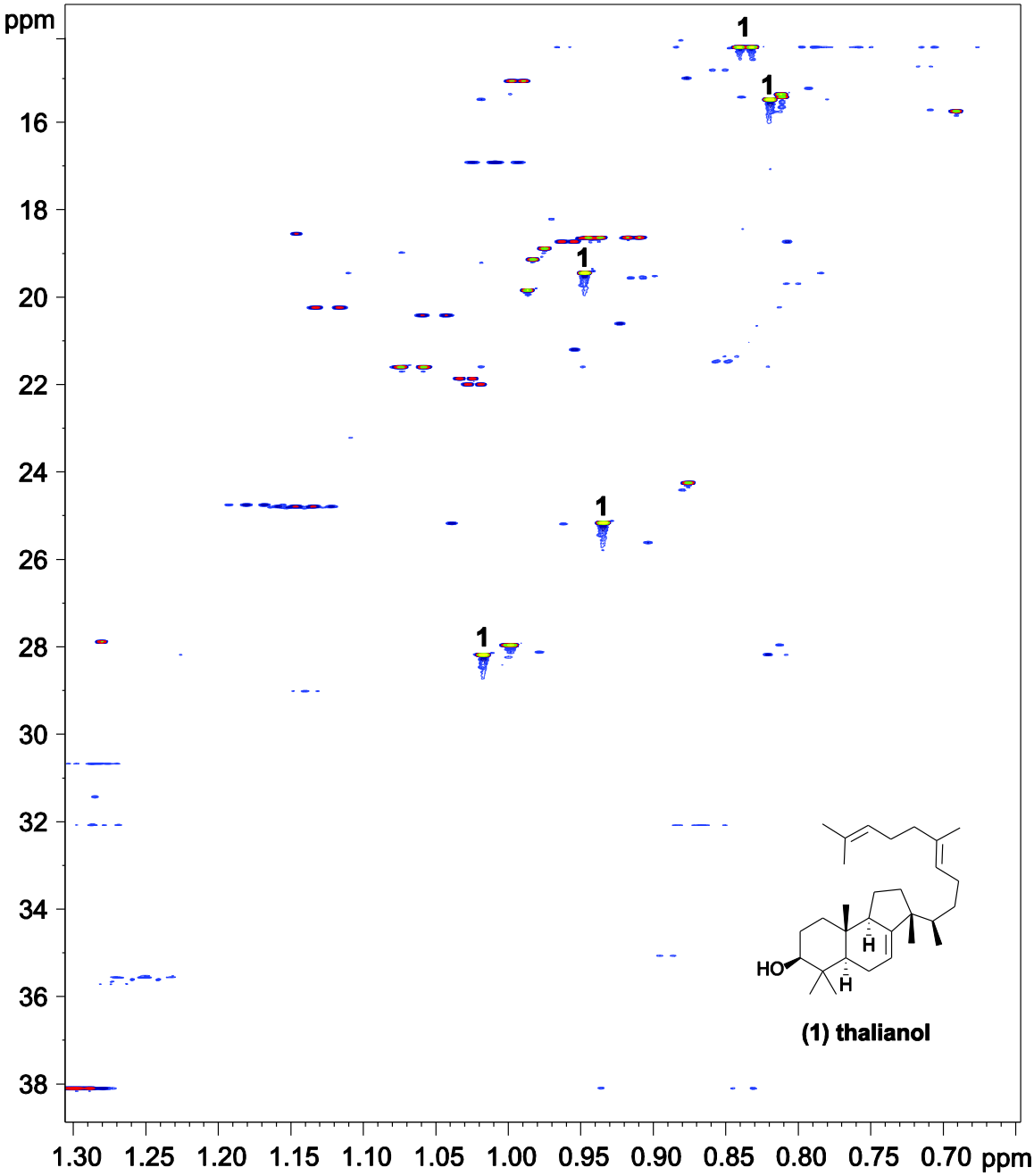


Figure E1.1. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vivo Aly*PEN4 SPE Fractions 22-35. Thalianol (**1**) represents the major product.

E1.2 HSQC spectra of *in vivo Aly*PEN6 SPE Fractions 21-35.

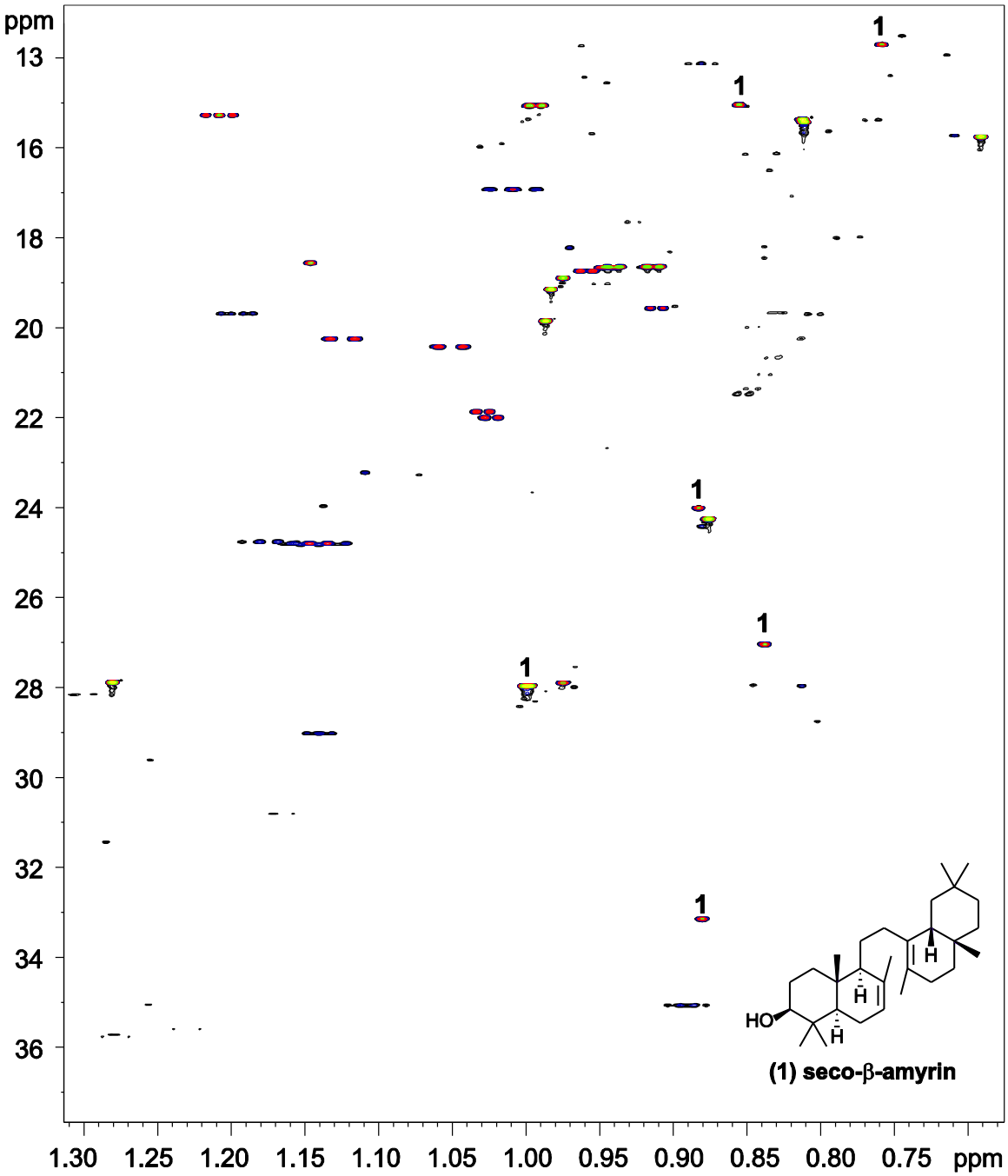


Figure E1.2. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vivo Aly*PEN6 SPE Fractions 21-35. Seco β-amyrin (**1**) represents the major product.

E1.3 HSQC spectra of *in vivo Aly*PEN9 SPE Fractions 23-30.

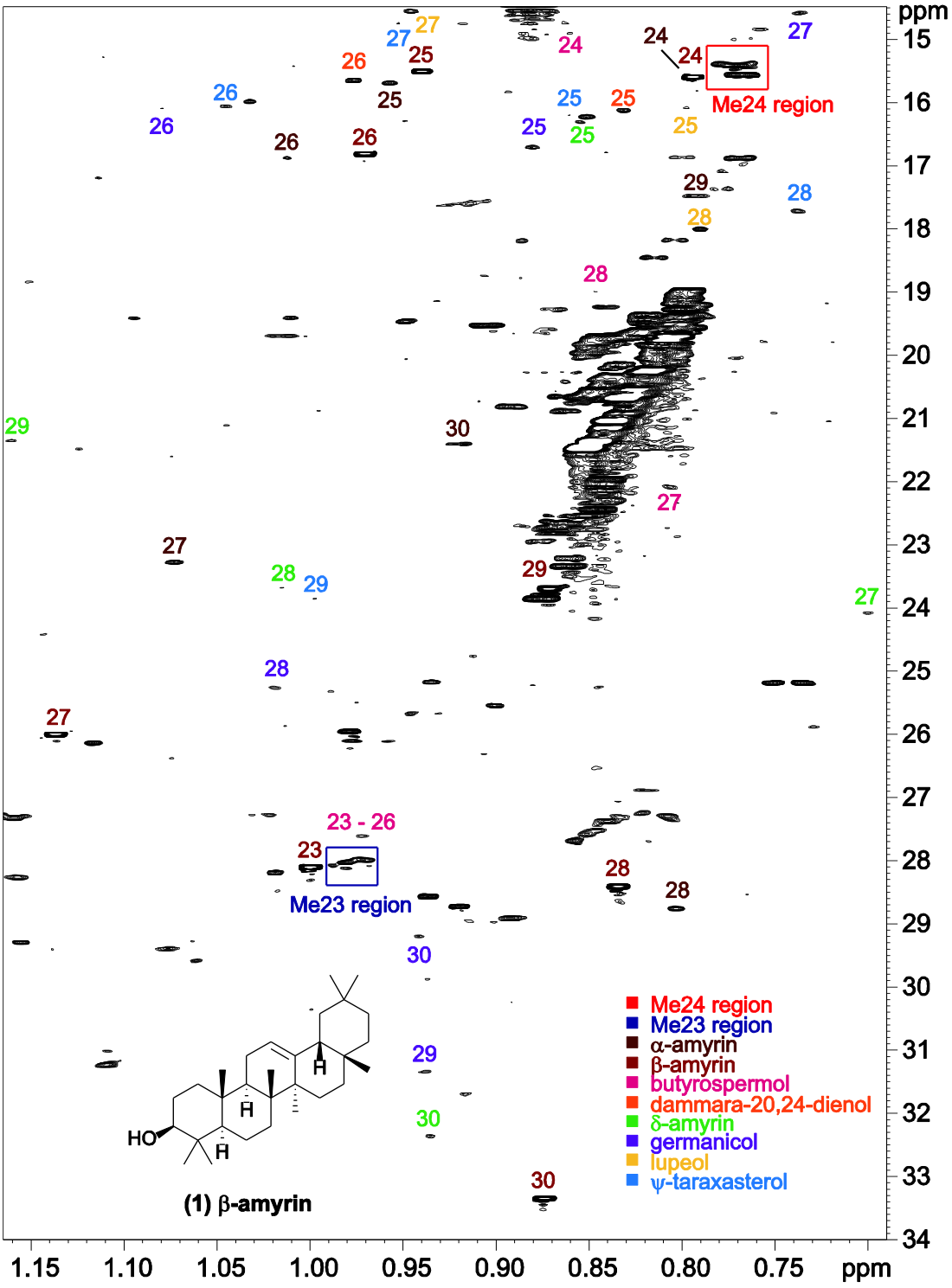


Figure E1.3. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vivo Aly*PEN9 SPE Fractions 23-30. β-amyrin (**1**) represents the major product.

E1.4 HSQC spectra of *in vivo L. perennis* nematocyphol synthase SPE Fractions 18-34.

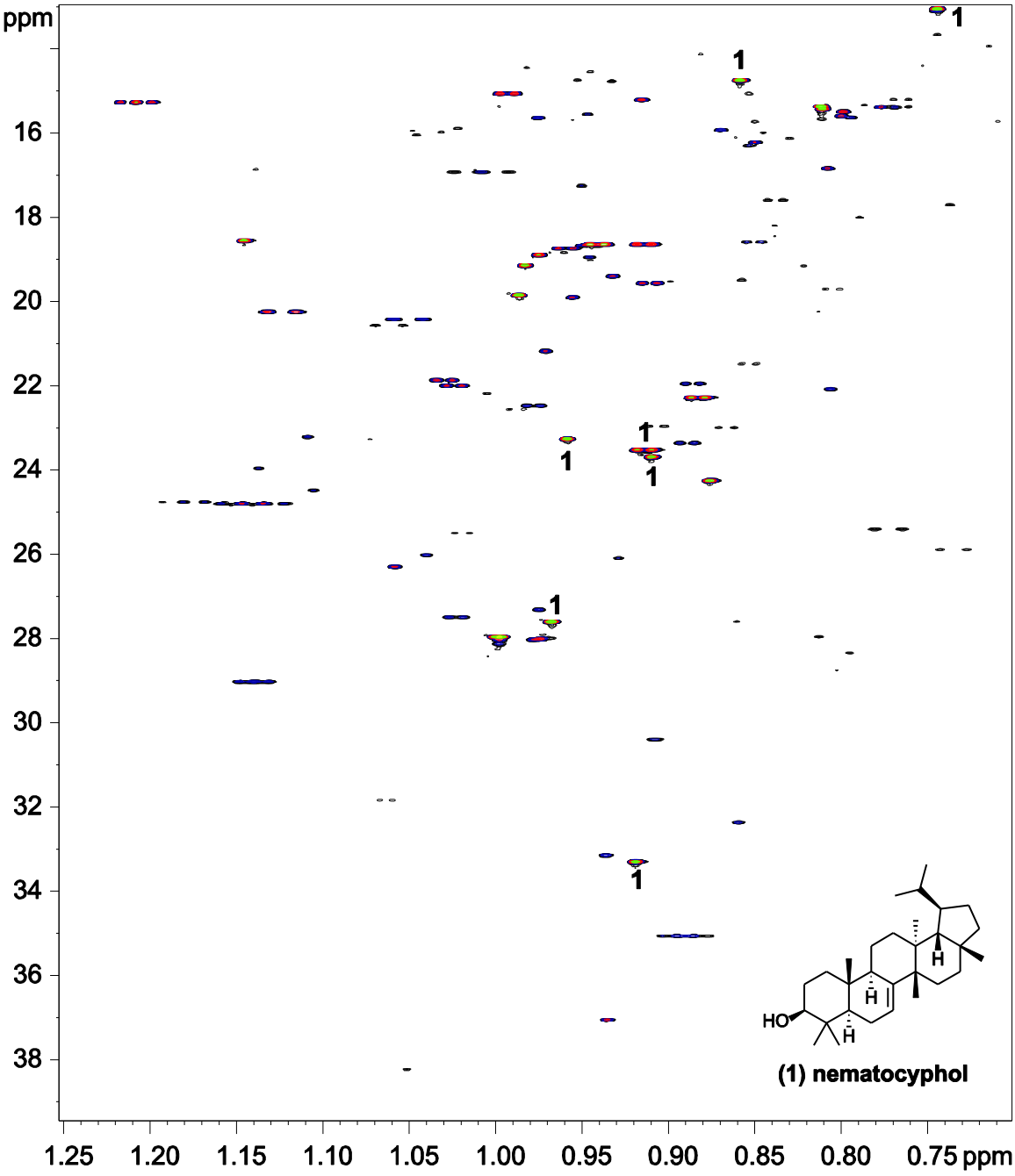


Figure E1.4. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vivo L. perennis* nematocyphol synthase SPE Fractions 18-34. Nematocyphol (**1**) represents the major product.

E1.5 HSQC spectra of *in vitro L. sativa* mixed amyrin synthase SPE Fractions 21-33.

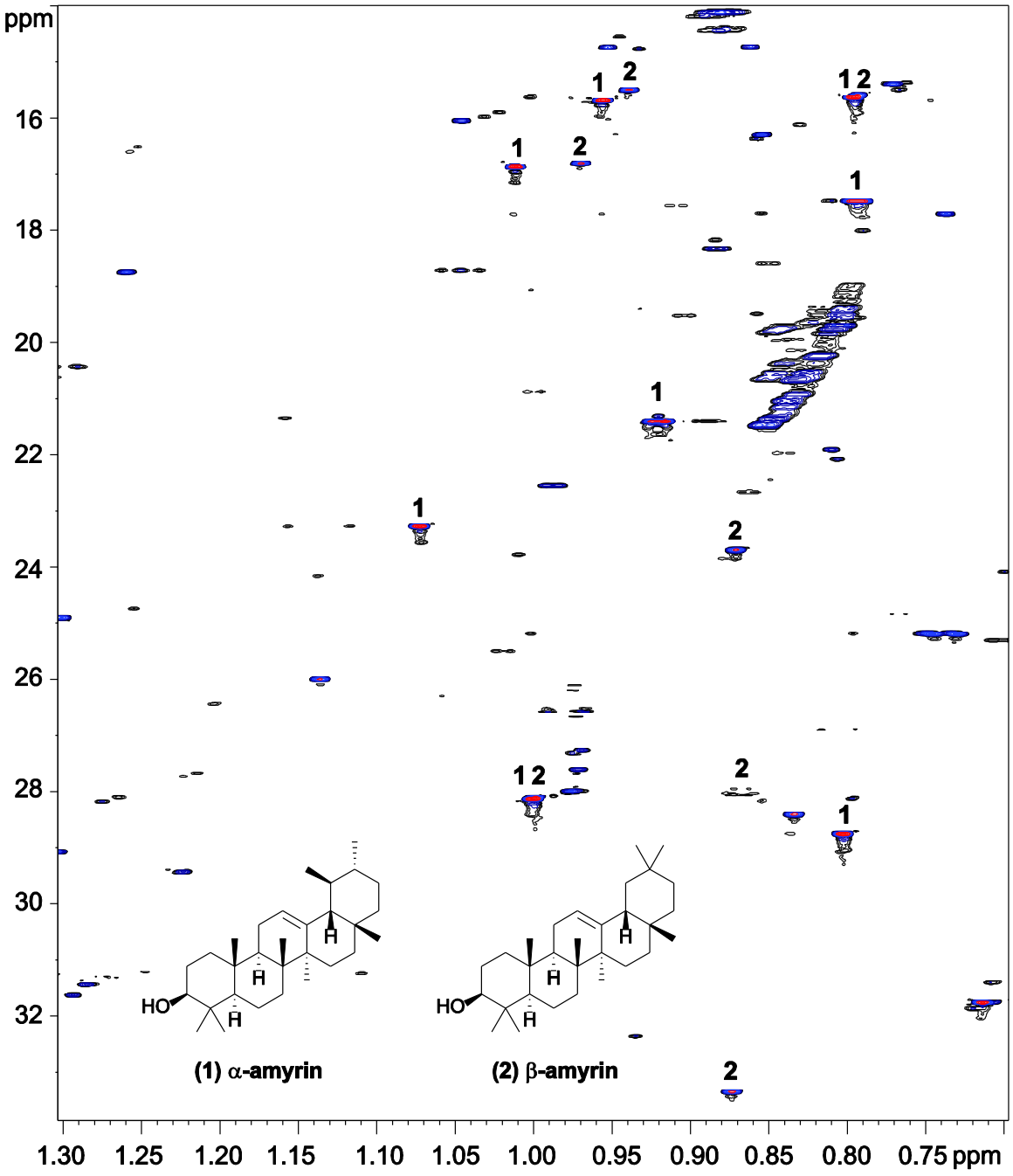


Figure E1.5. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vitro L. sativa* α-amyrin synthase SPE Fractions 21-33. α-amyrin (**1**) and (**2**) β-amyrin represent the two major product.

E1.6 HSQC spectra of *in vitro L. serriola* mixed taraxasterol synthase SPE Fractions 22-32.

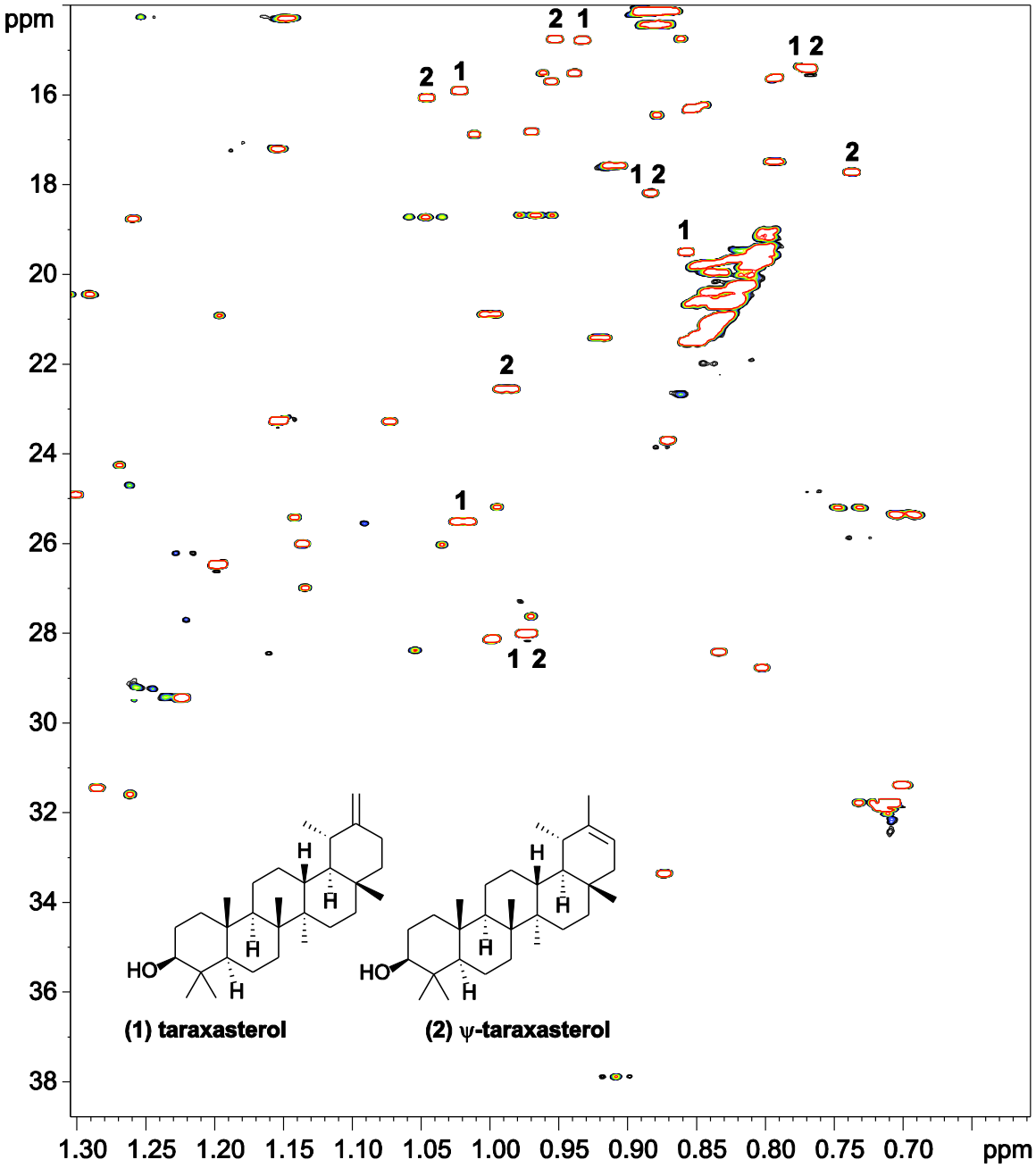


Figure E1.6. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vitro L. serriola* mixed taraxasterol synthase SPE Fractions 22-32. Taraxasterol (**1**) and (**2**) Ψ-taraxasterol represent the two major product.

E1.7 HSQC spectra of *in vitro L. serriola* lupeol synthase SPE Fractions 21-32.

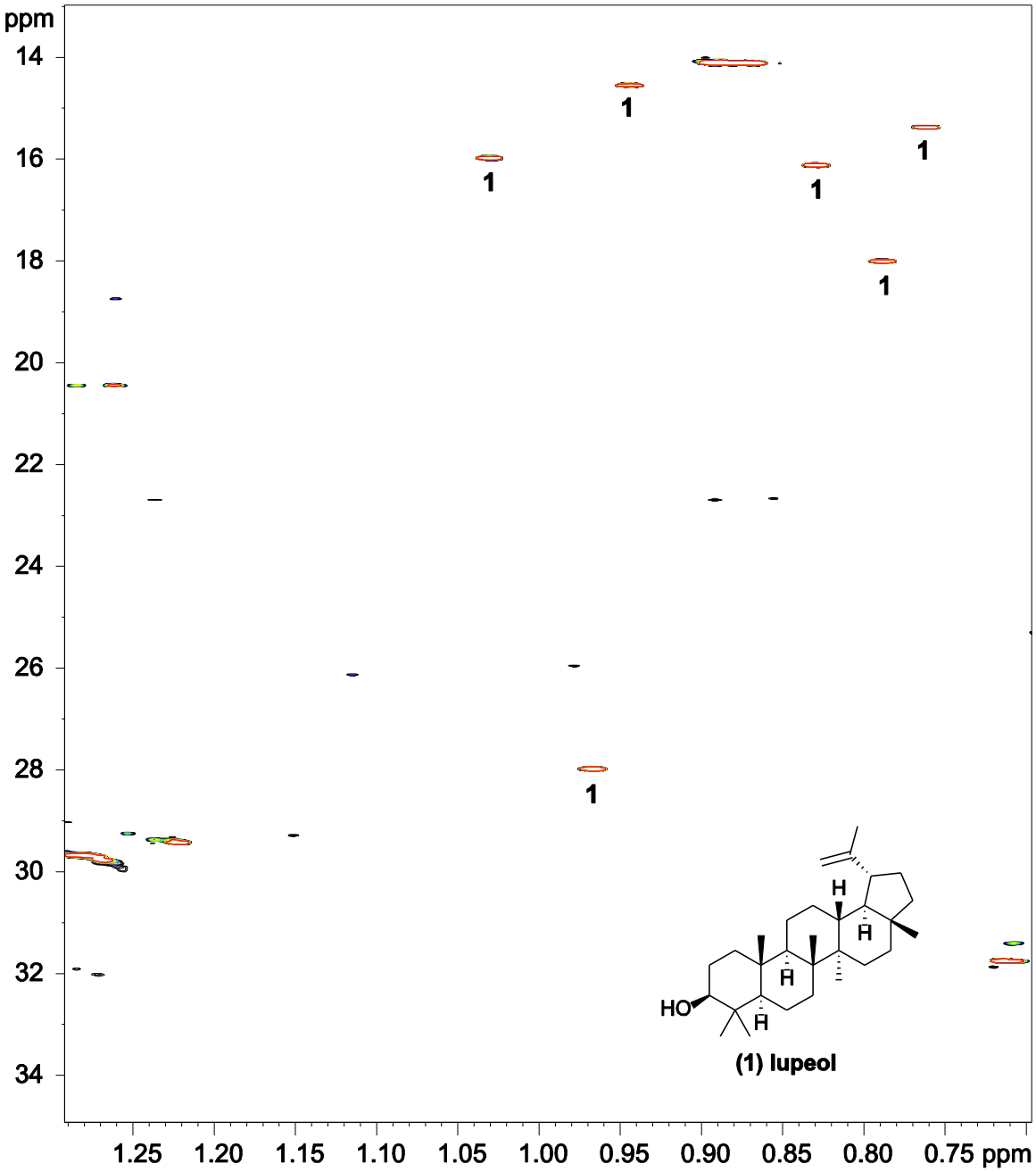


Figure E1.7. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vitro L. serriola* lupeol synthase SPE Fractions 21-32. Lupeol (**1**) represents the major product.

E1.8 HSQC spectra of *in vitro L. serriola* cycloartenol synthase SPE Fractions 21-32.

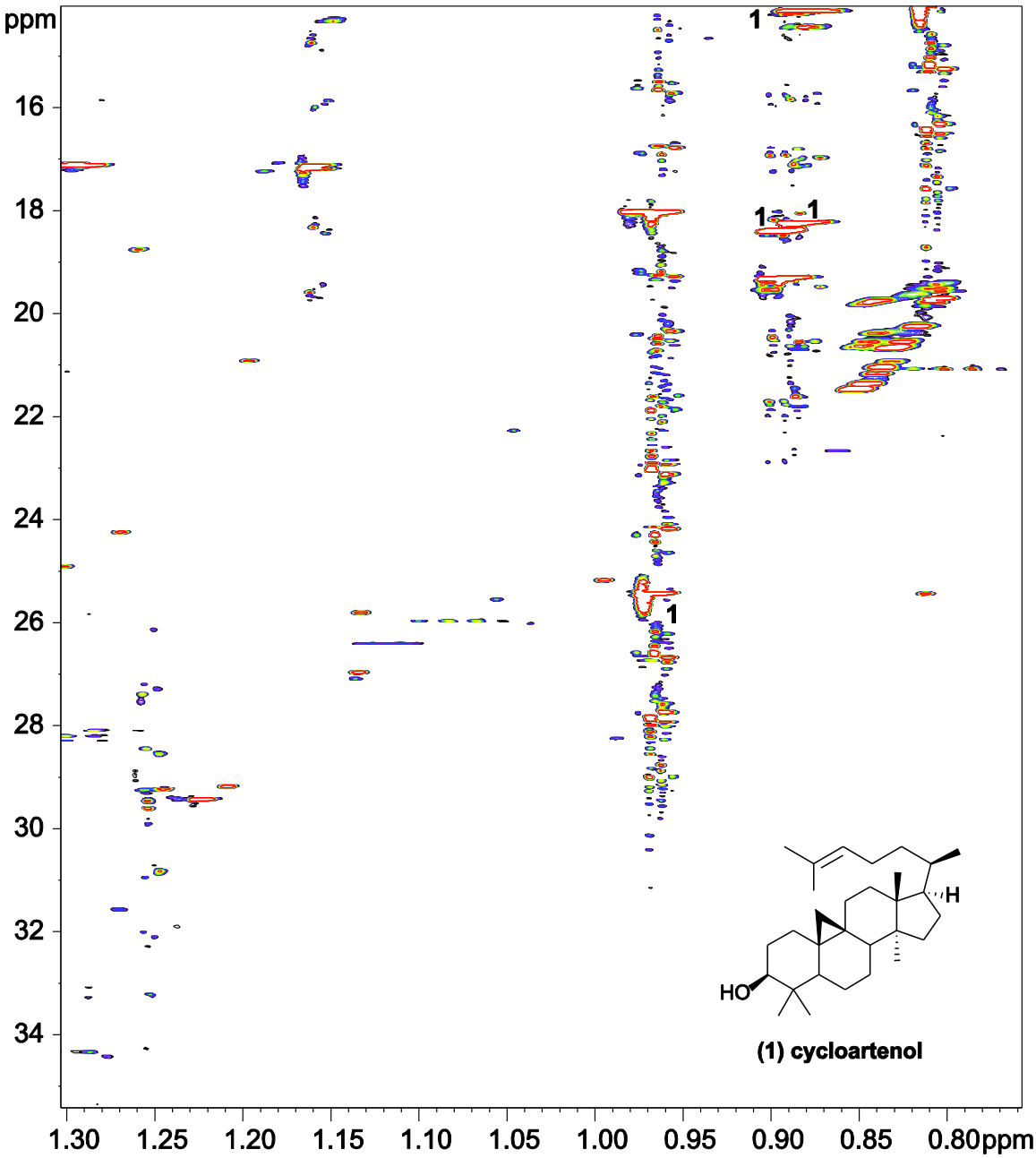


Figure E1.8. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vitro L. serriola* cycloartenol synthase SPE Fractions 21-32. Cycloartenol (**1**) represents the major product.

E1.9 HSQC spectra of *in vivo L. serriola* tirucalla-8,24-dien-3β-ol.

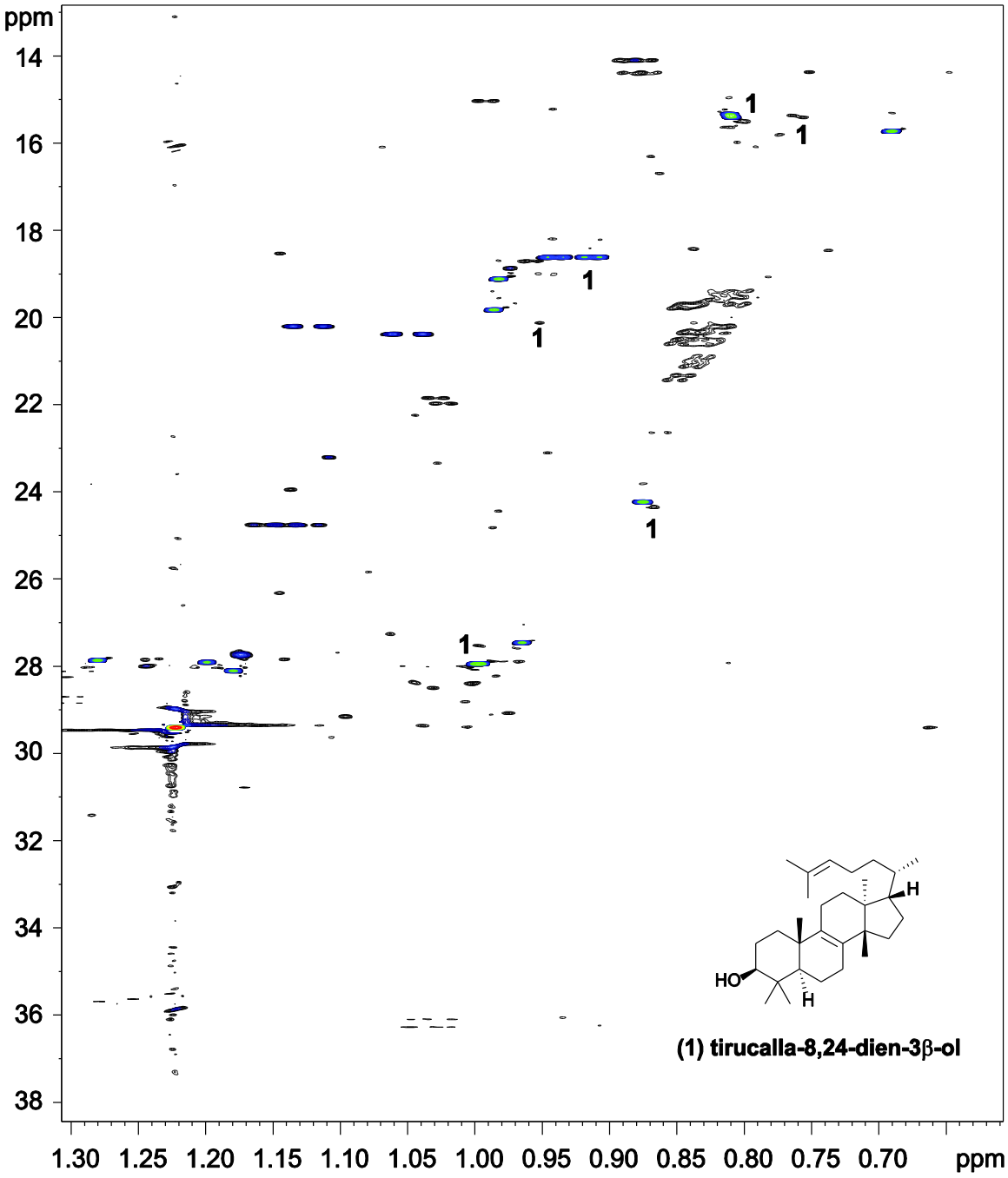


Figure E1.9. Upfield methyl region of the 800 MHz H and HSQC spectra of *in vivo L. serriola* tirucalla-8,24-dienol synthase SPE Fractions 4-9. Tirucalla-8,24-dien-3β-ol (**1**) represents the major product.

E1.10 HSQC spectra of crude extract Lettuce flowers.

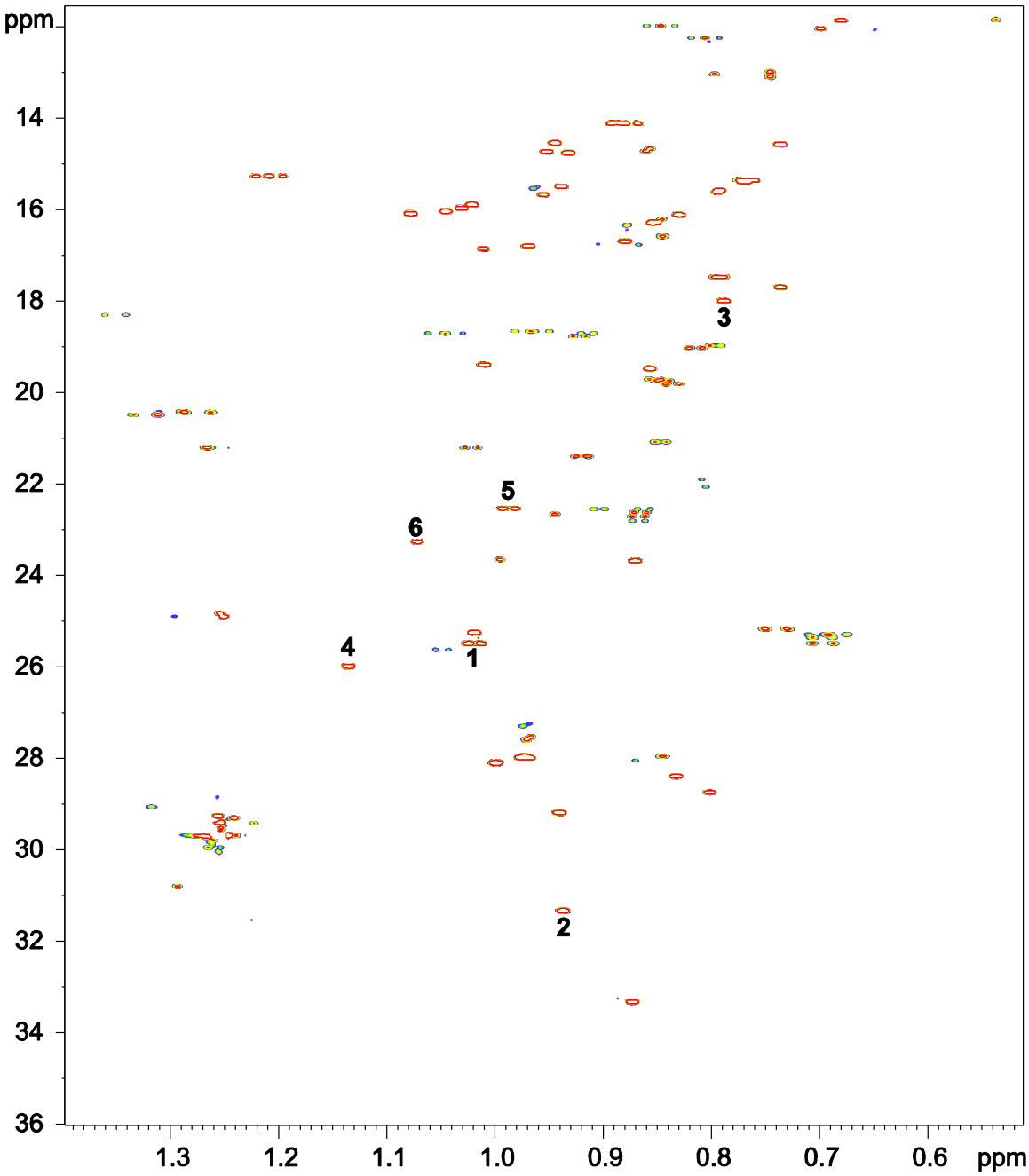


Figure E1.10. Upfield methyl region of the 800 MHz H and HSQC spectra of the crude extract of Lettuce flowers. Taraxasterol (**1**), germanicol (**2**), lupeol (**3**), β-amyrin (**4**), Ψ-taraxasterol (**5**), α-amyrin (**6**) represent the major products.

E1.11 HSQC spectra of crude extract Lettuce roots.

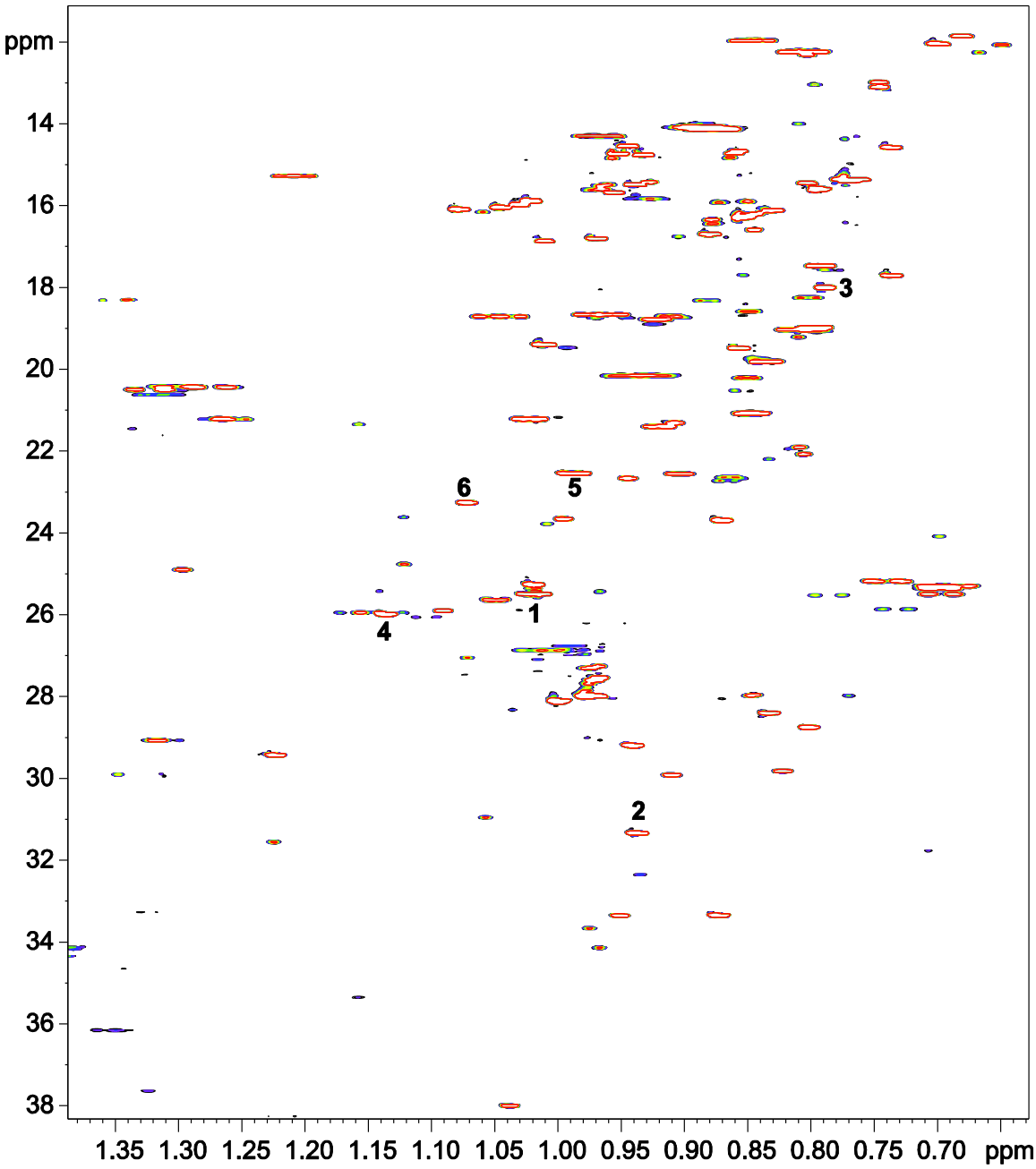


Figure E1.11. Upfield methyl region of the 800 MHz H and HSQC spectra of the crude extract of Lettuce roots. Taraxasterol (**1**), germanicol (**2**), lupeol (**3**), β-amyrin (**4**), Ψ-taraxasterol (**5**), α-amyrin (**6**) represent the major products.

E1.12 HSQC spectra of crude extract Lettuce leaves.

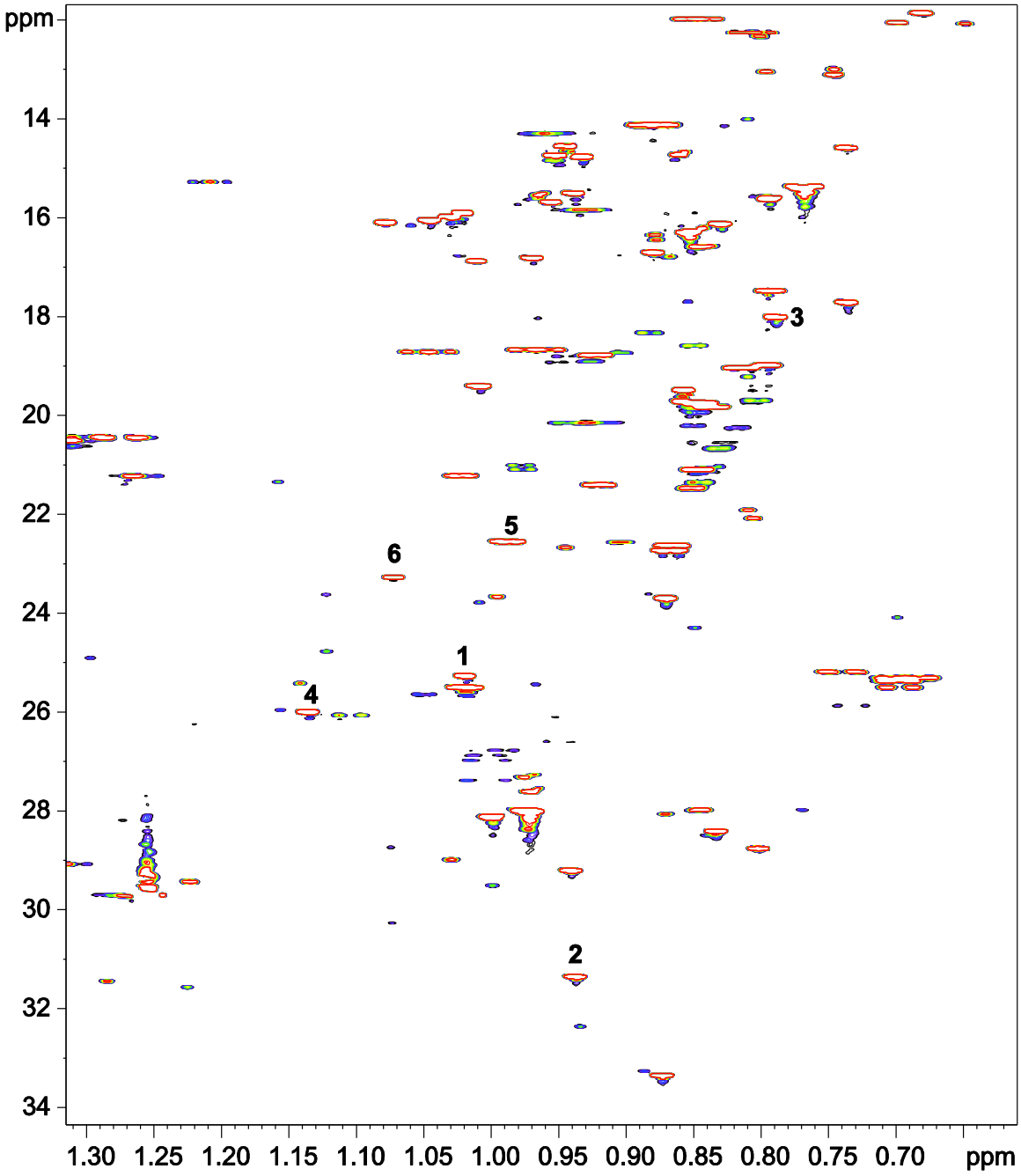


Figure E1.12. Upfield methyl region of the 800 MHz H and HSQC spectra of the crude extract of Lettuce leaves. Taraxasterol (**1**), germanicol (**2**), lupeol (**3**), β-amyrin (**4**), Ψ-taraxasterol (**5**), α-amyrin (**6**) represent the major products.

E1.13 HSQC spectra of crude extract Lettuce stem.

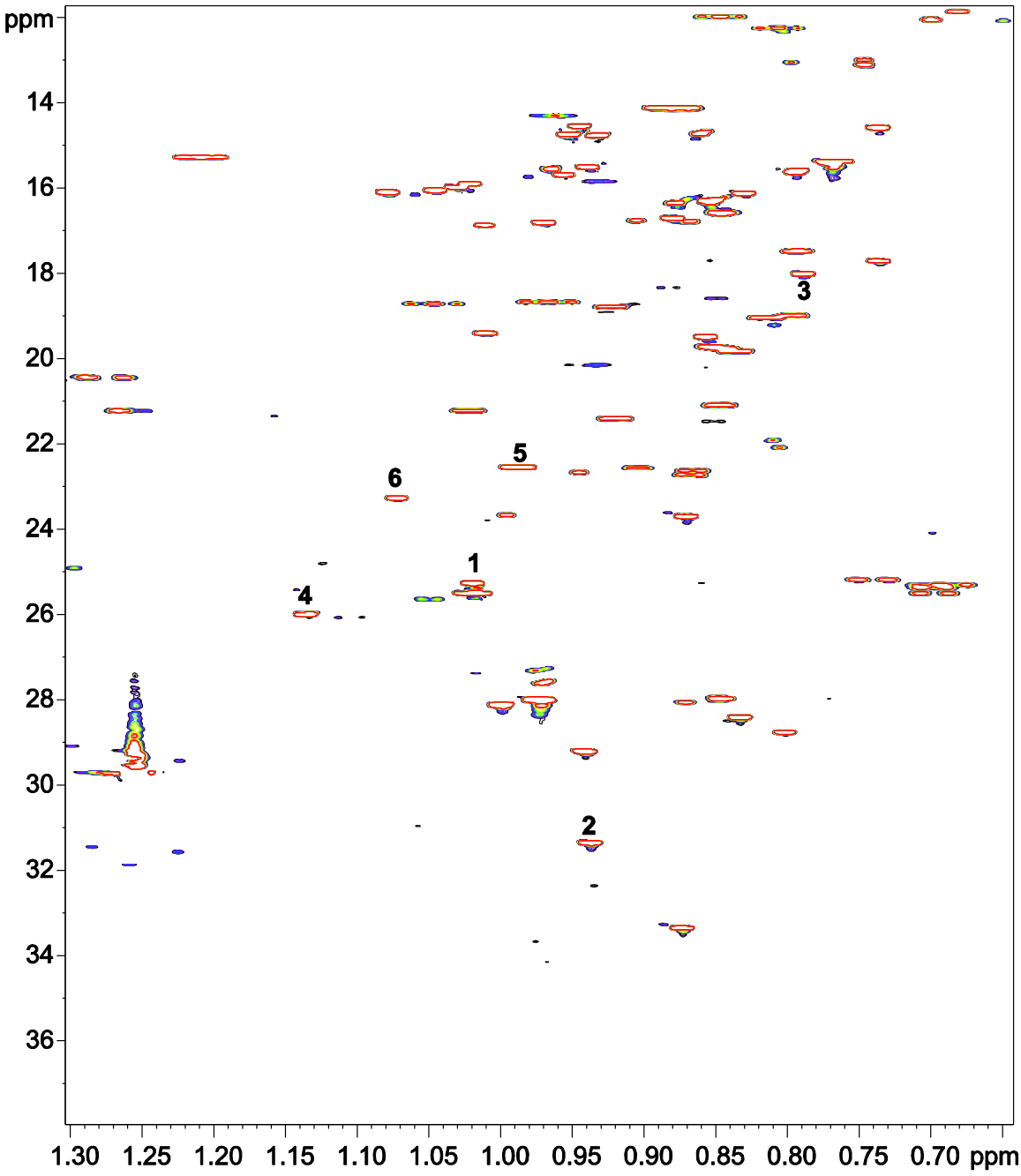


Figure E1.13. Upfield methyl region of the 800 MHz H and HSQC spectra of the crude extract of Lettuce stem. Taraxasterol (**1**), germanicol (**2**), lupeol (**3**), β-amyrin (**4**), Ψ-taraxasterol (**5**), α-amyrin (**6**) represent the major products.

E1.14 HSQC spectra of crude extract Lettuce seeds.

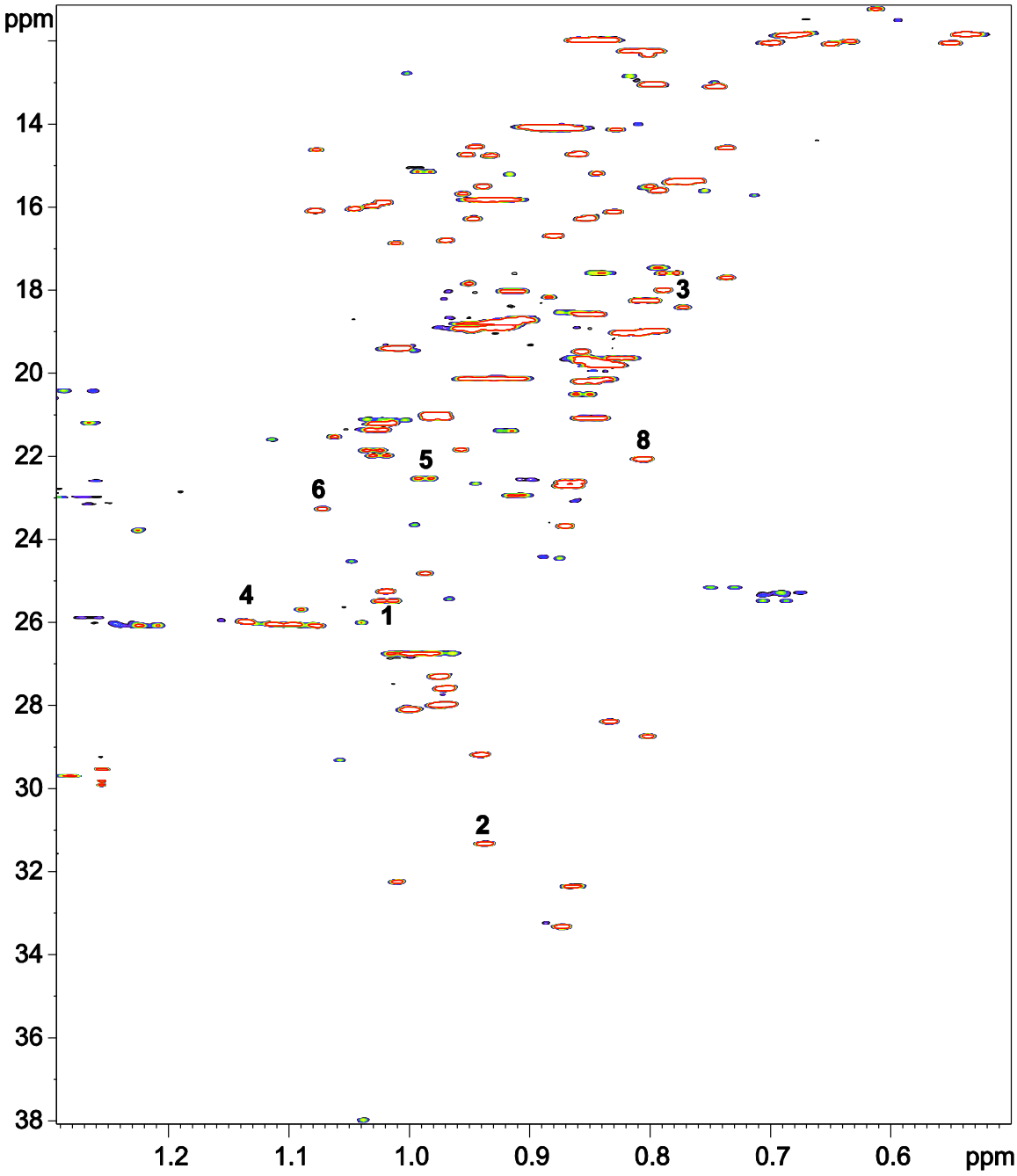


Figure E1.14. Upfield methyl region of the 800 MHz H and HSQC spectra of the crude extract of Lettuce seeds. Taraxasterol (**1**), germanicol (**2**), lupeol (**3**), β-amyrin (**4**), Ψ-taraxasterol (**5**), α-amyrin (**6**) and butyrospermol (**8**) represent the major products.

*** Compound numbering for Figures E1.10 to E1.14 comes from Chapter 5 Table 5.12.**

E1.15 HSQC spectra of crude extract *Arabidopsis lyrata* seeds.

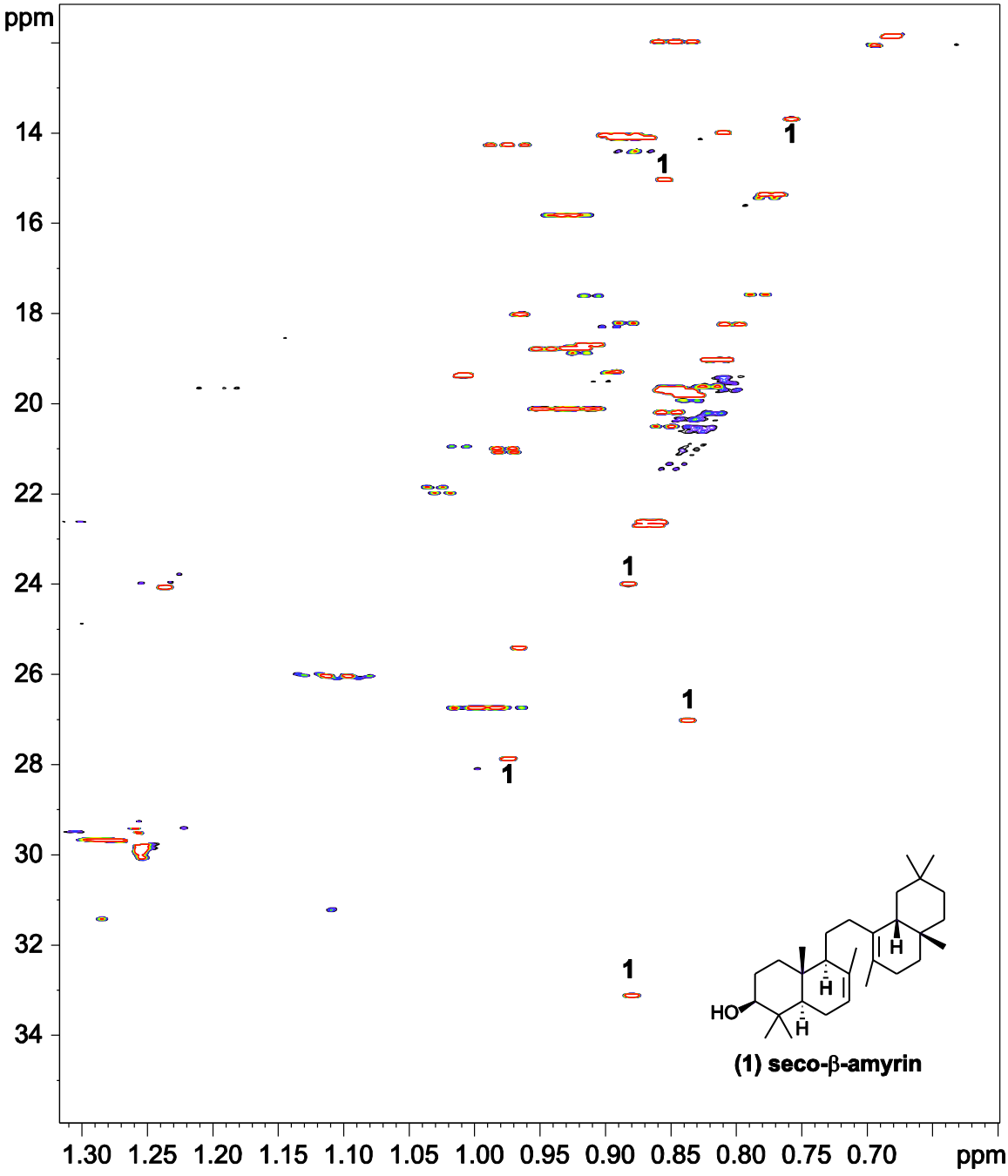


Figure E1.15. Upfield methyl region of the 800 MHz H and HSQC spectra of the crude extract of *Arabidopsis lyrata* seeds. Seco β-amyrin (**1**) represents the major product.

Appendix F

F1. GC-MS Spectral Data

Table F1 letter code assigned to the OSCs or plant extract studied in this work.

A. *Aly*PEN4
B. *Aly*PEN6
C. *Aly*PEN9
D. *A. lyrata* seeds extract
E. *L. perennis* nematocyphol synthase
F. *L. sativa* mixed amyrin synthase
G. *L. serriola* mixed taraxasterol synthase
H. *L. serriola* lupeol synthase
I. *L. serriola* cycloartenol synthase
J. *L. serriola* tirucalla-8,24-dienol synthase
K. Lettuce plant extract

Table F1. Summary of the triterpene alcohols found in this work. Numbers in red represent the number that was assigned to a specific compound when analyzing a particular OSC or plant extract.

#	Product name	A	B	C	D	E	F	G	H	I	J	K
F1.1	(20*R*)-dammar-24-ene-3β,20-diol		19	10		32	9	7				
F1.1	(20*S*)-dammar-24-ene-3β,20-diol	10	13	7	13	20	6	3				
F1.2	(20*R/S*)-taraxastane-3β,20-diol		40			35	20					35
F1.3	Δ7-14-epithalianol	5	36									
F1.3	Δ7-thalianol	20	24			36						
F1.3	14-epithalianol	4	38									
F1.4	Δ8(26)-polypodatetraenol-9βH											23
F1.5	Δ8(26)-seco-β-amyrin		7		7	38						
F.16	17-epilupeol					24						
F.17	18-lupane-3b-ol					17						
F1.8	25-nor-9β-methylpodioda-5(10),17E,21-trien-3β-ol	3										
F1.9	achilleol A		18									24
F1.10	arabidiol	14										
F1.10	14-epiarabidiol	2										
F1.11	bauerenol		6		6	21	15	13				7

F1.12	butyrospermol		20	9		5	7	6	3			8
F1.13	camelliol C	8	16									
F1.14	curcurbitadienol									4		16
F1.15	cycloartenol									1		12
F1.16	dammara-20(22)*E*,24-dien-3β-ol	18	23			34						
F1.17	dammara-20(22)*Z*,24-dien-3β-ol		29			37						
F1.18	dammara-20,24-dien-3β-ol	16	9	2	9	4	3	17	9			
F1.19	DB-friedo-lup-5-en-3β-ol					19						
F1.19	lup-19(21)-en-3β-ol					13						
F1.20	euphol				23							20
F1.21	germanicol	21	31	6			18		5			2
F1.22	glutinol		15		15							
F1.23	isobauerenol		37			11	14	16				
F1.24	isoeuphol		32			28						
F1.25	isotirucallol		33			29	16	14			3	
F1.26	isoursenol		17			2	12	15				17
F1.27	lanosterol									3		
F1.27	parkeol									2		18
F1.28	lupane-3β,20-diol		11		11	18	19	11	2			10
F1.29	lupeol		2	3	2	12	10	8	1			3
F1.30	multiflorenol		14		14		21					13
F1.31	nematocyphol		22			1	13	12	8			14
F1.32	neolup-13(18)-enol					25						
F1.33	neolupenol					7						
F1.34	polypoda-7,13*E*,17*E*,21-tetraen-3β-ol		3									
F1.35	seco-α-amyrin		12		12	10						
F1.36	seco-β-amyrin		1		1	31						
F1.37	taraxasterol	22	27	11		8	8	1	10			19
F1.38	taraxerol		28			27	17					19
F1.39	tarolupenol					6						21
F1.40	thalianol	1										
F1.41	tirucalla-7,24-dien-3β-ol	20	35			15	5	9	7		2	
F1.42	tirucalla-8,24-dien-3β-ol	9				33			6		1	
F1.43	tylolupenol A					3						
F1.44	α-amyrin	24	4	4	4	14	1	5				6
F1.45	β-amyrin	23	30	1		26		4	4			4

F1.46	γ-amyrin		34			16						
F1.47	δ-amyrin		8	8	8	22	11	10				11
F1.48	ψ-taraxasterol	25	10	5	10	9	4	2				5
**	13αH-malabarica-14(27)-trien-3b-ol	12										
**	13βH-malabarica-14(27)-trien-3b-ol	15	26									
**	13αH-malabarica-14*E*,17*E*,21-trien-3b-ol	7	25									
**	13αH-malabarica-14*Z*,17*E*,21-trien-3b-ol	6										
**	13βH-malabarica-14*Z*,17*E*,21-trien-3β-ol	27	21									
**	13βH-malabarica-14*E*,17*E*,21-trien-3b-ol	17										
**	20*S*-dammara-12,24-dienol										4	
**	boeticol											22
**	friedelin		39									
**	polypoda-8,13*E*,17*E*,21-tetraen-3β-ol	13										
**	9αH-polypoda-7,13*E*,17*E*,21-tetraen-3β-ol	11			3							
**	9αH-polypoda-8(26),13*E*,17*E*,21-tetraen-3β-ol	19	5		5							
**	rhoiptenol					30						

**Found by HSQC NMR.

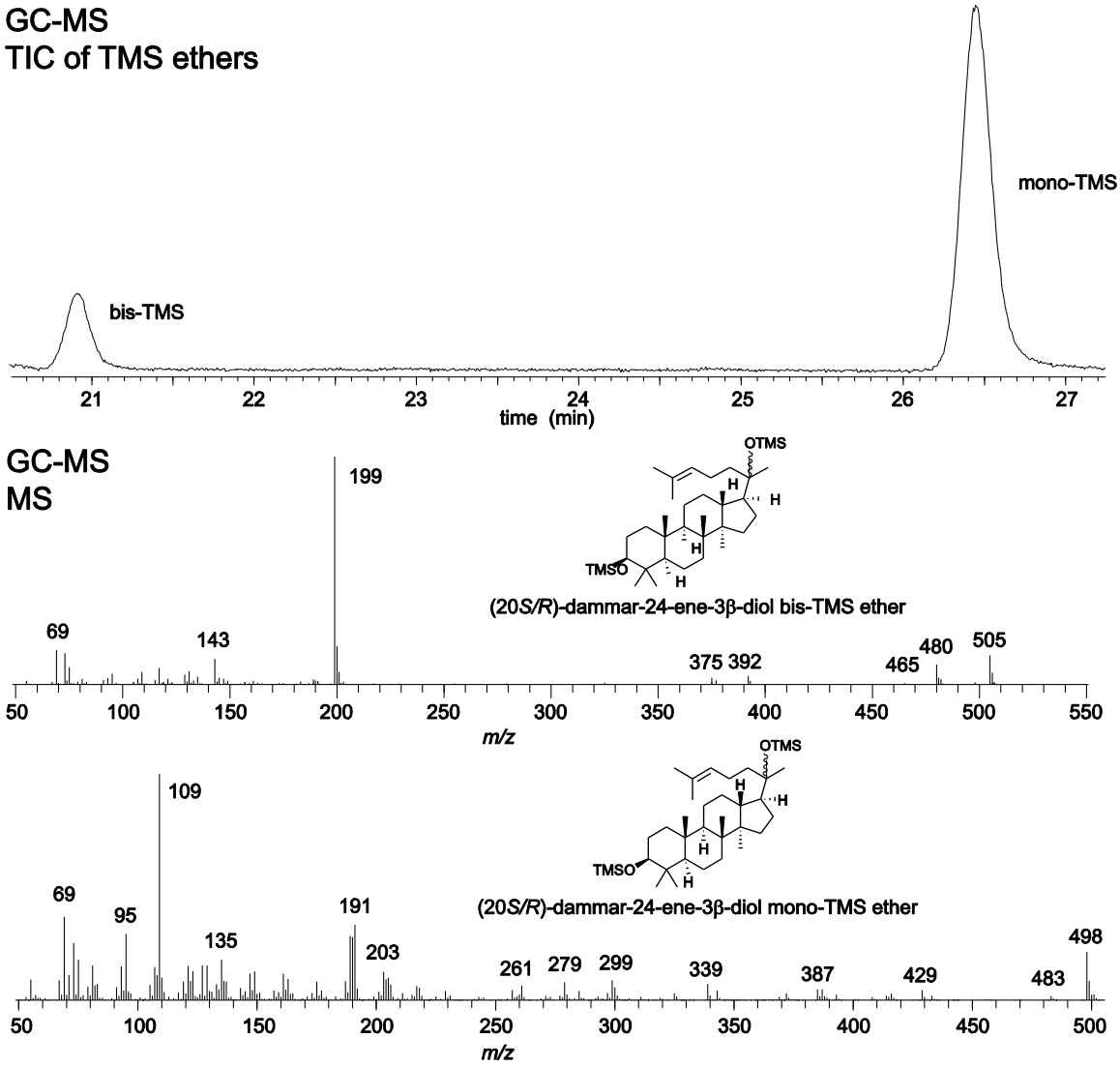


Figure F1.1. GC-MS TIC and EI mass spectra of (20*S*/*R*)-dammar-24-ene-3β,20-diol.

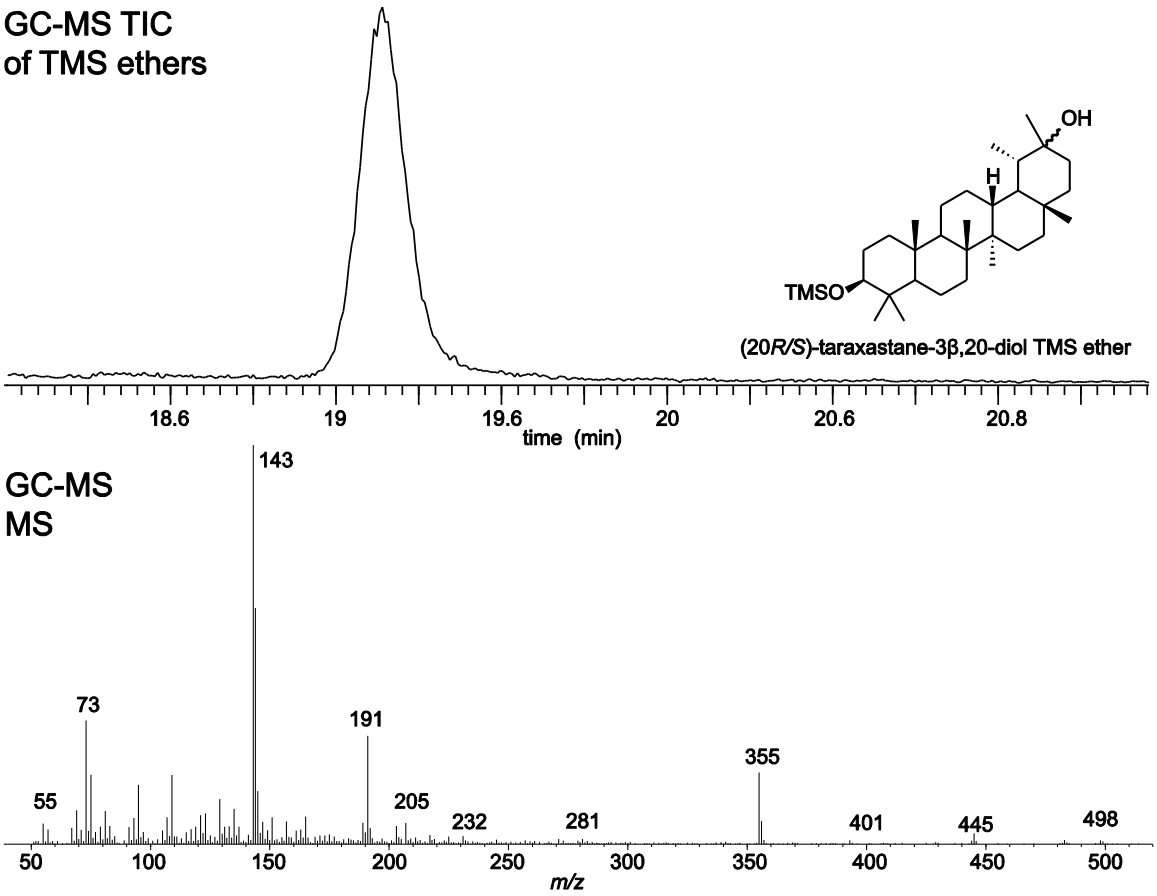


Figure F1.2. GC-MS TIC and EI mass spectra of (20*R/S*)-taraxastane-3β,20-diol.

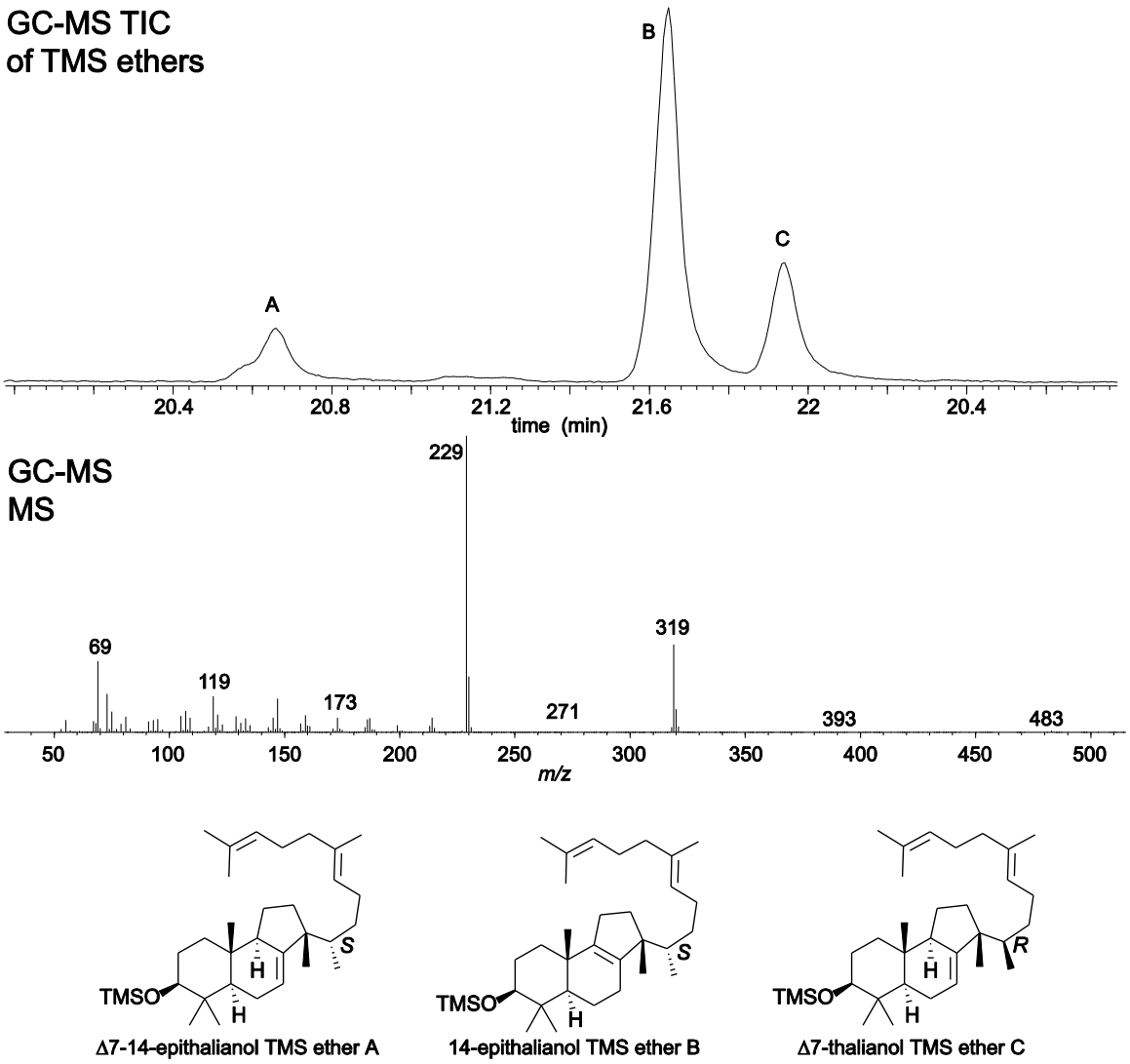


Figure F1.3. GC-MS TIC and EI mass spectra of Δ7-14-epithalianol, Δ7-thalianol and 14-epithalianol.

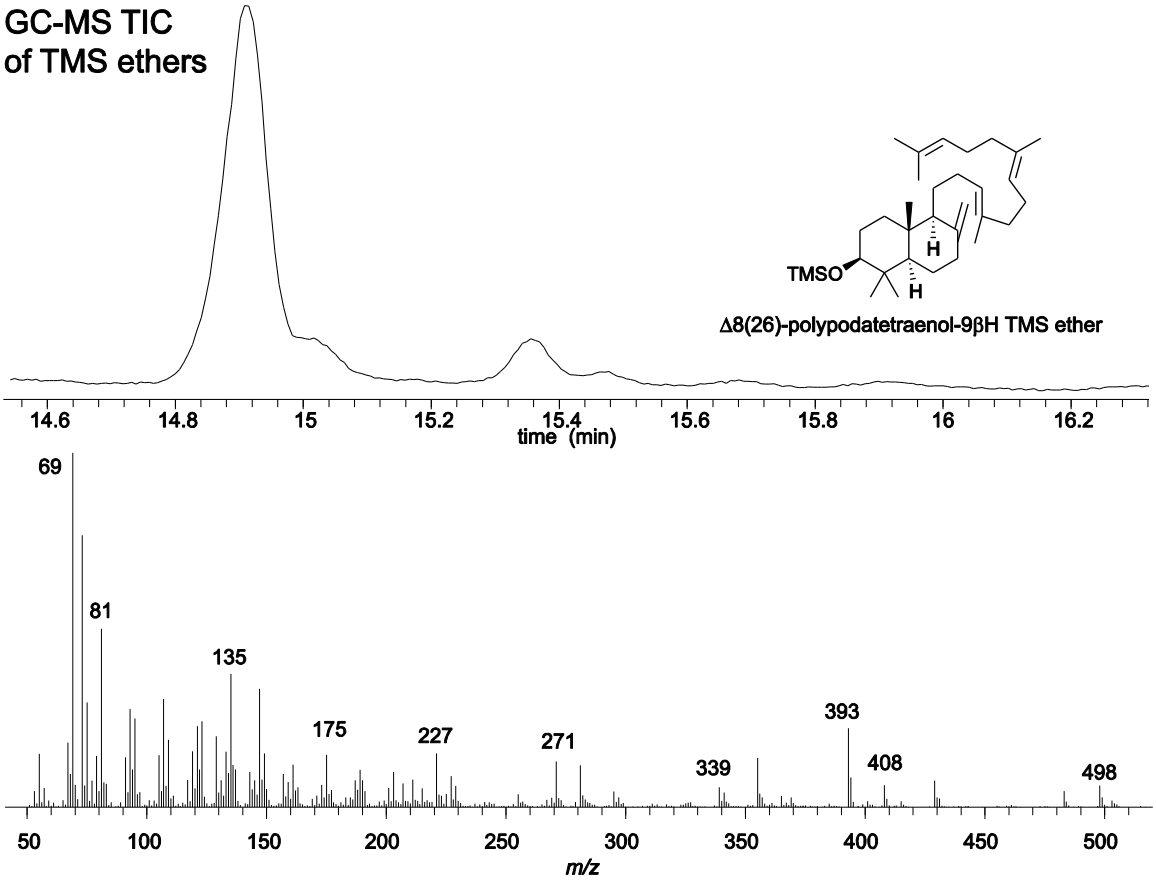


Figure F1.4. GC-MS TIC and EI mass spectra of Δ8(26)-polypodatetraenol-9βH.

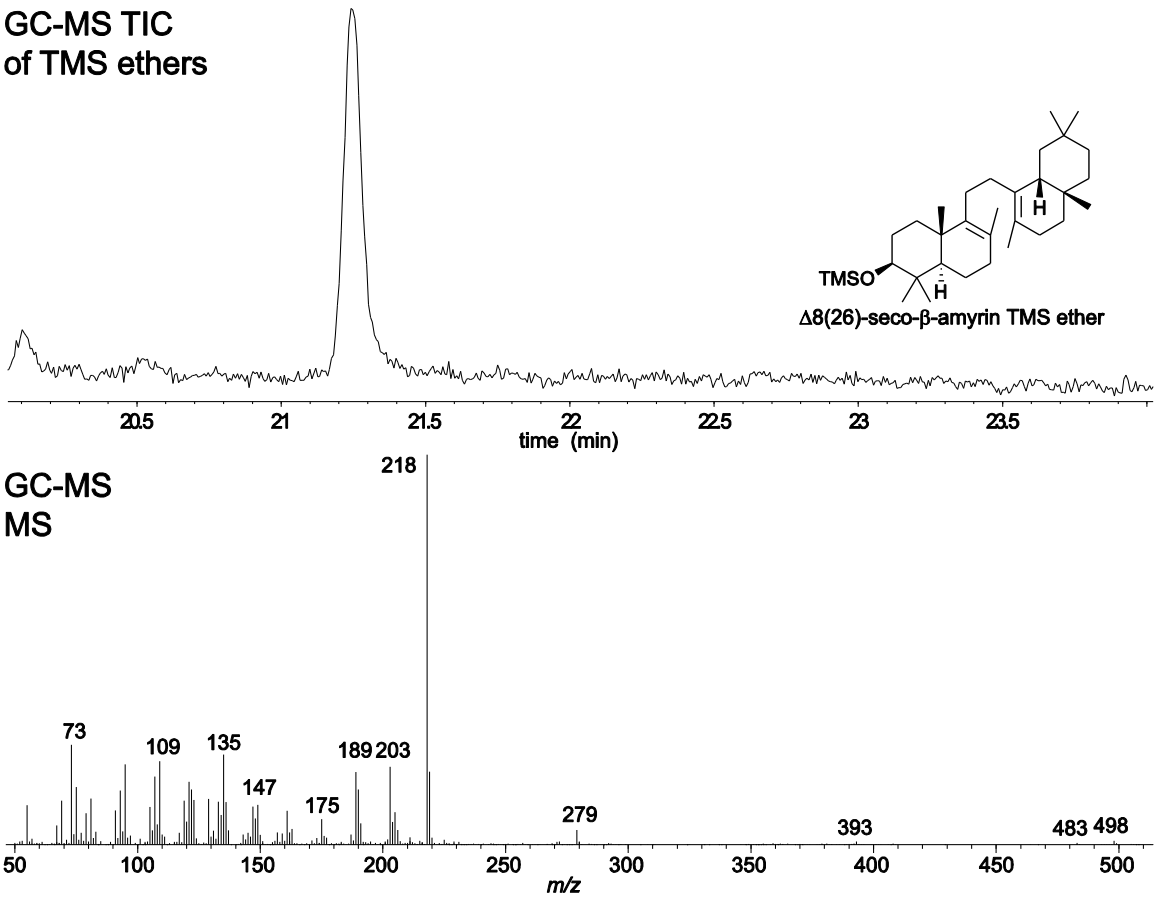


Figure F1.5. GC-MS TIC and EI mass spectra of Δ8(26)-seco-β-amyrin.

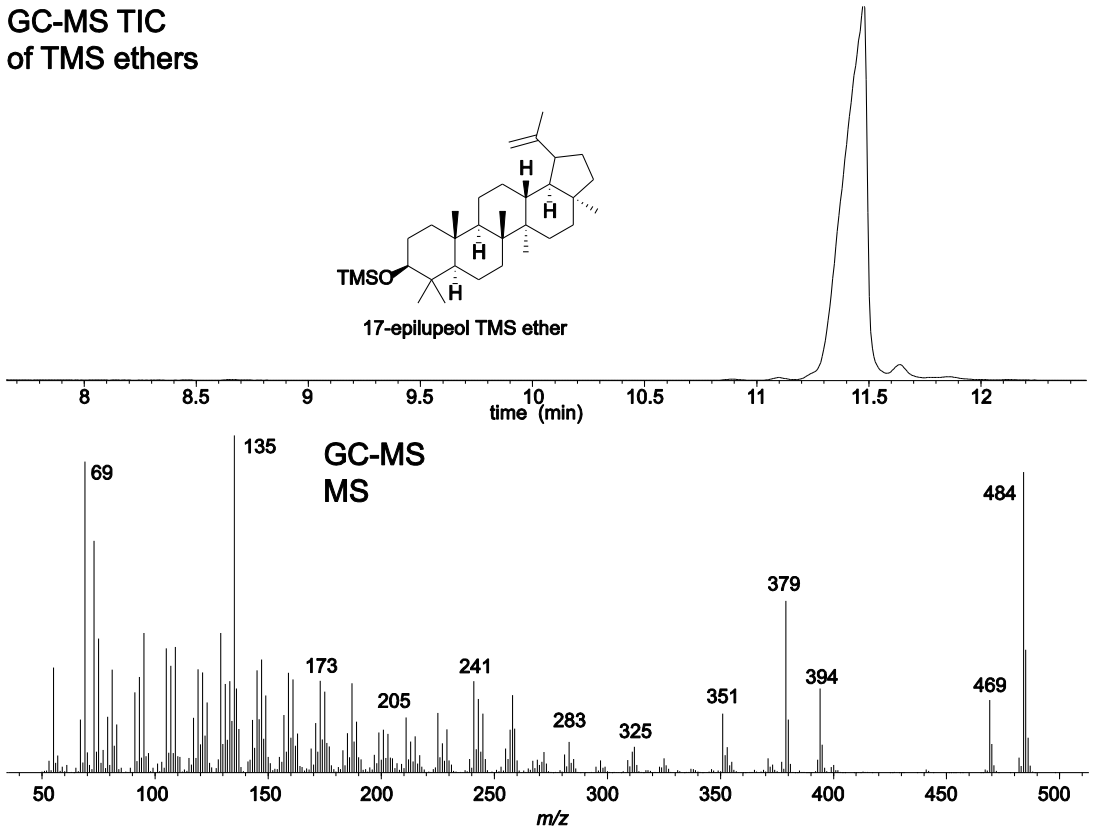


Figure F1.6. GC-MS TIC and EI mass spectra of 17-epilupeol.

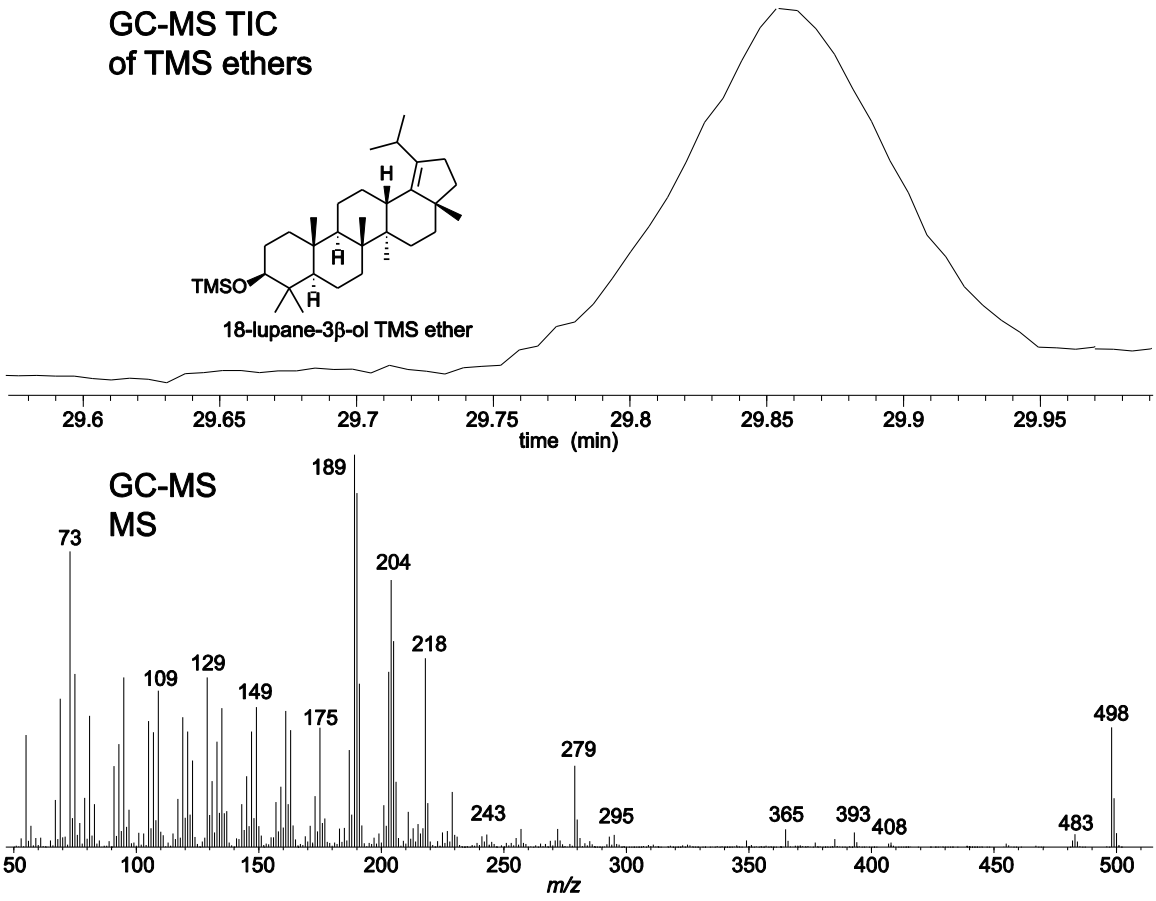


Figure F1.7. GC-MS TIC and EI mass spectra of 18-lupane-3b-ol.

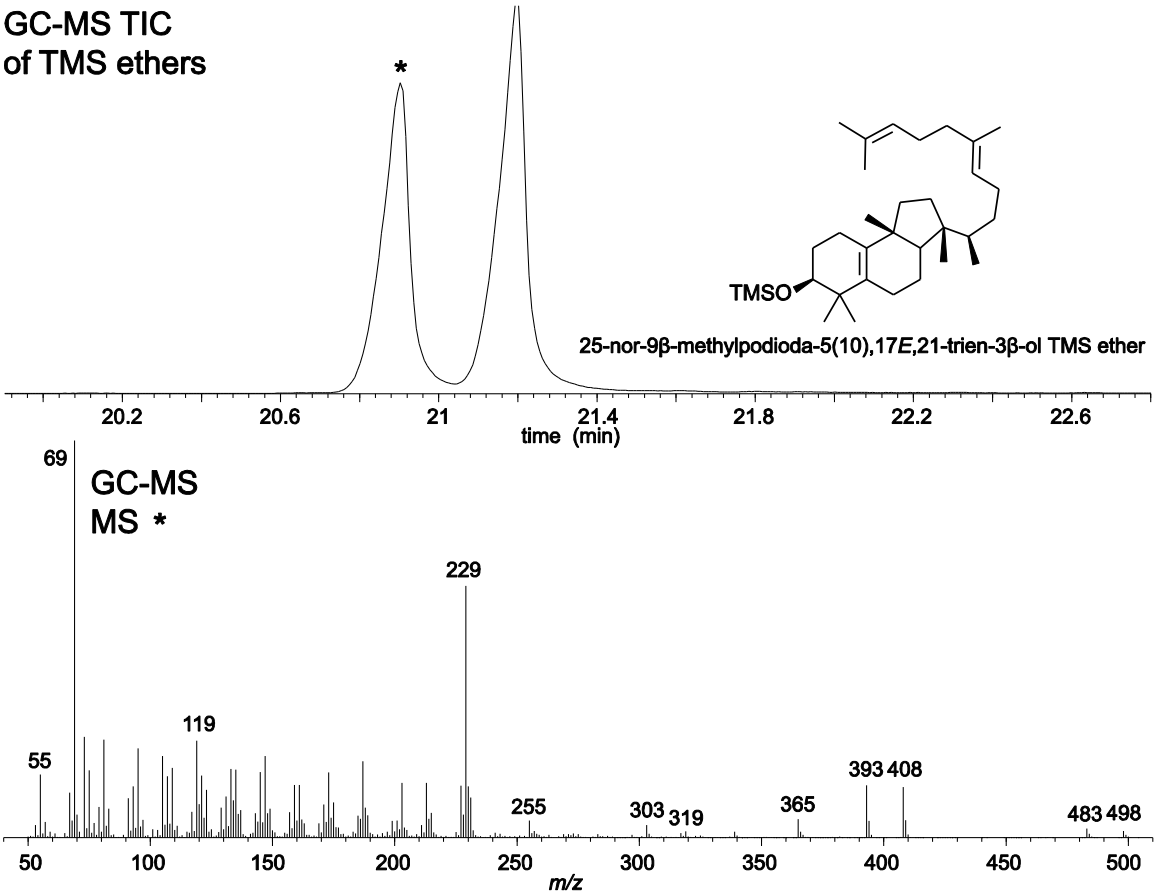


Figure F1.8. GC-MS TIC and EI mass spectra of 25-nor-9β-methylpodioda-5(10),17*E*,21-trien-3β-ol.

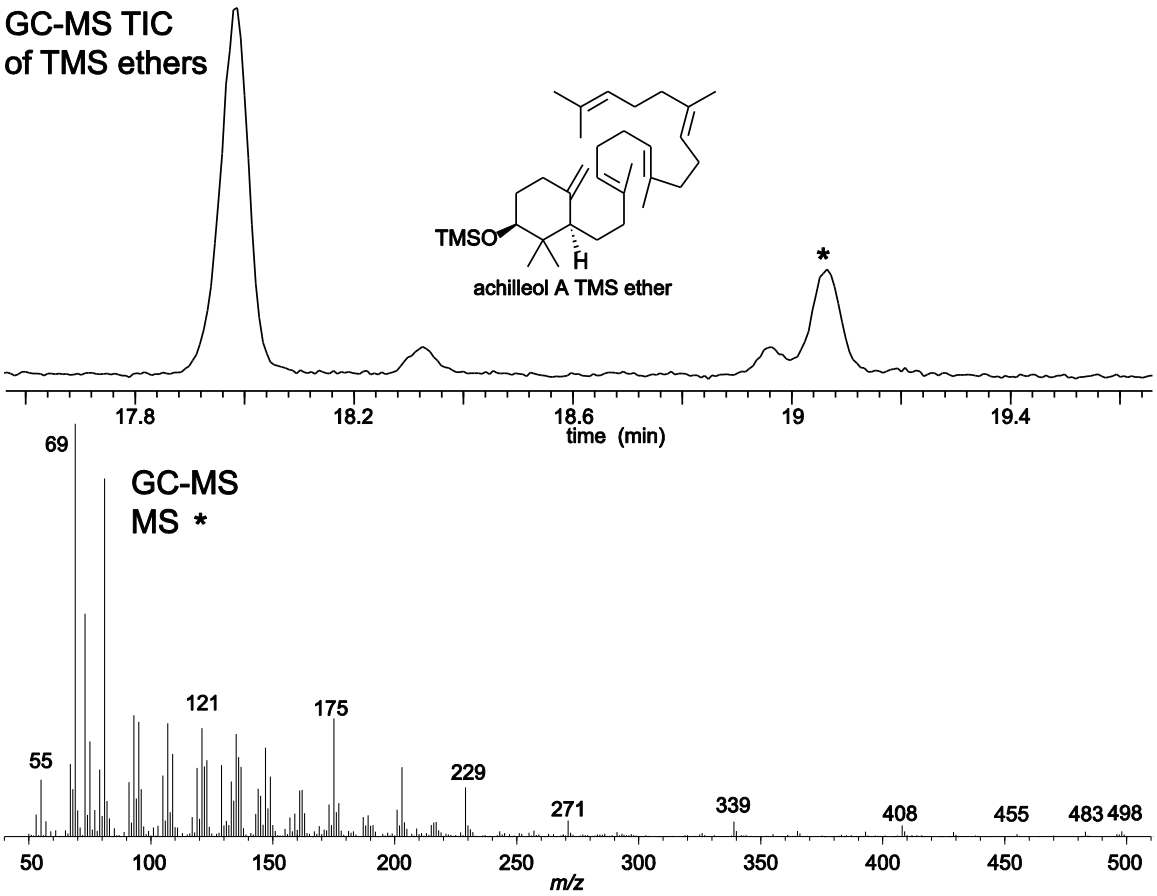


Figure F1.9. GC-MS TIC and EI mass spectra of achilleol A.

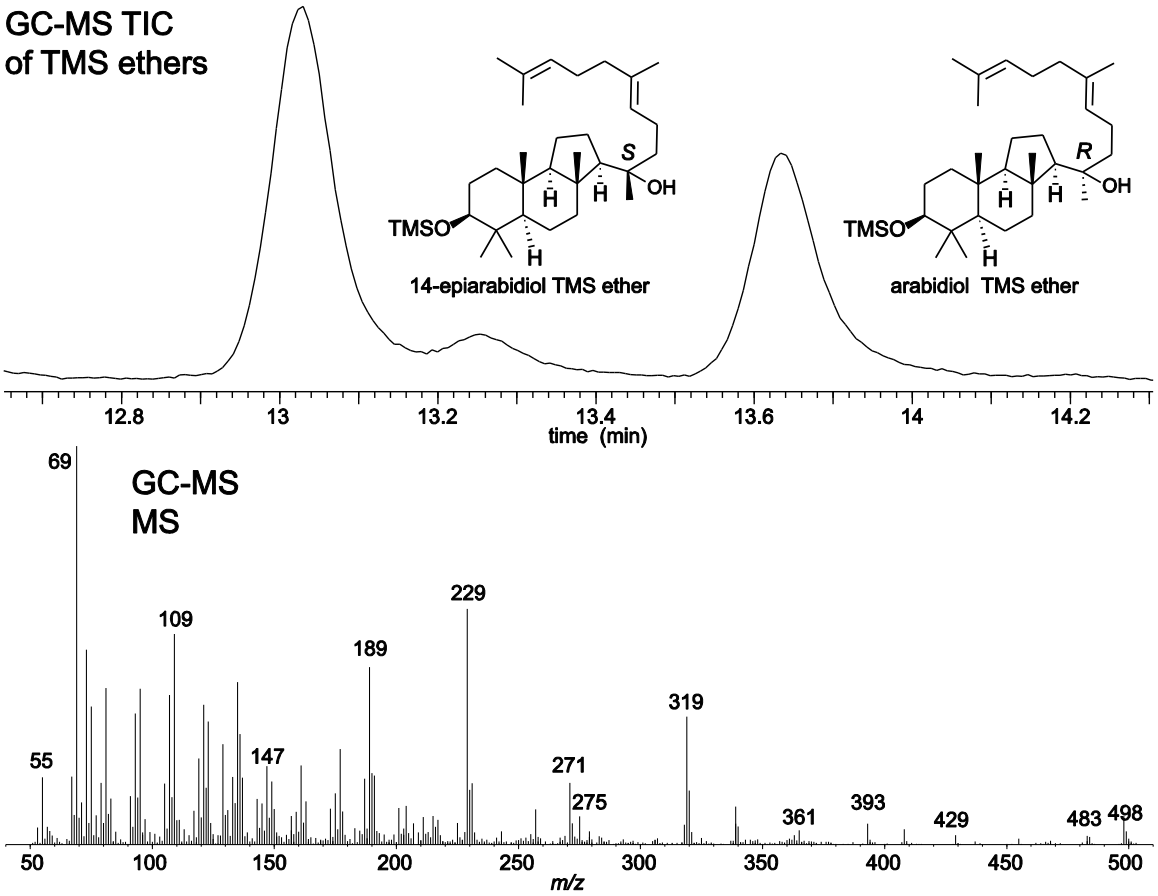


Figure F1.10. GC-MS TIC and EI mass spectra of arabidiol and 14-epiarabidiol.

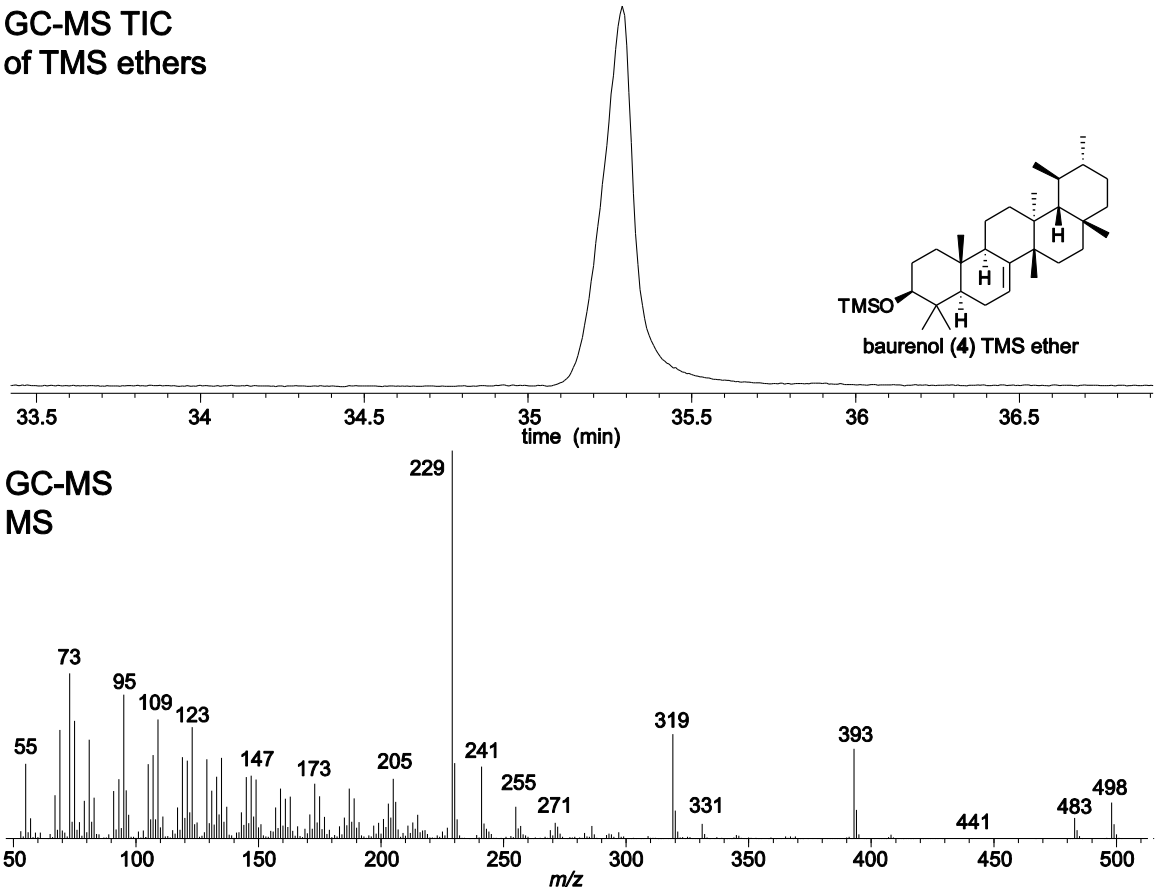


Figure F1.11. GC-MS TIC and EI mass spectra of bauerenol.

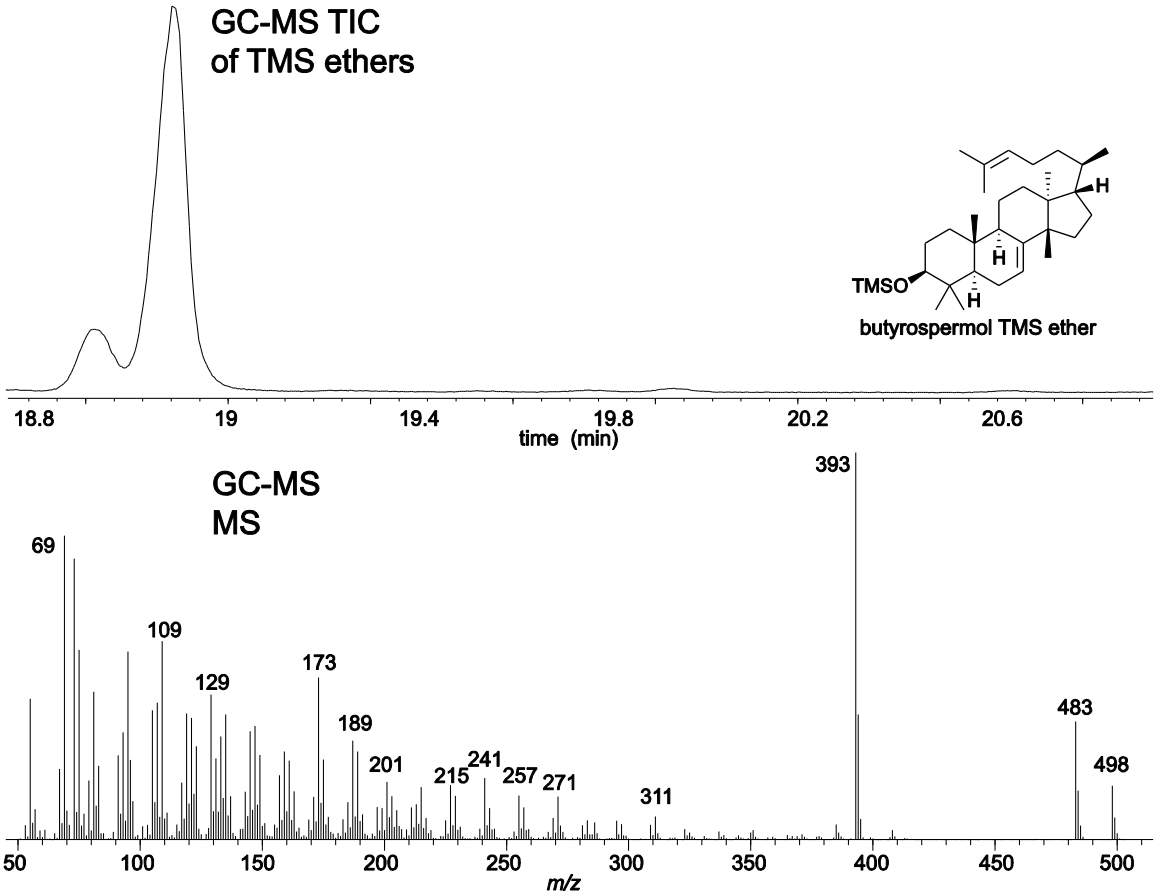


Figure F1.12. GC-MS TIC and EI mass spectra of butyrospermol.

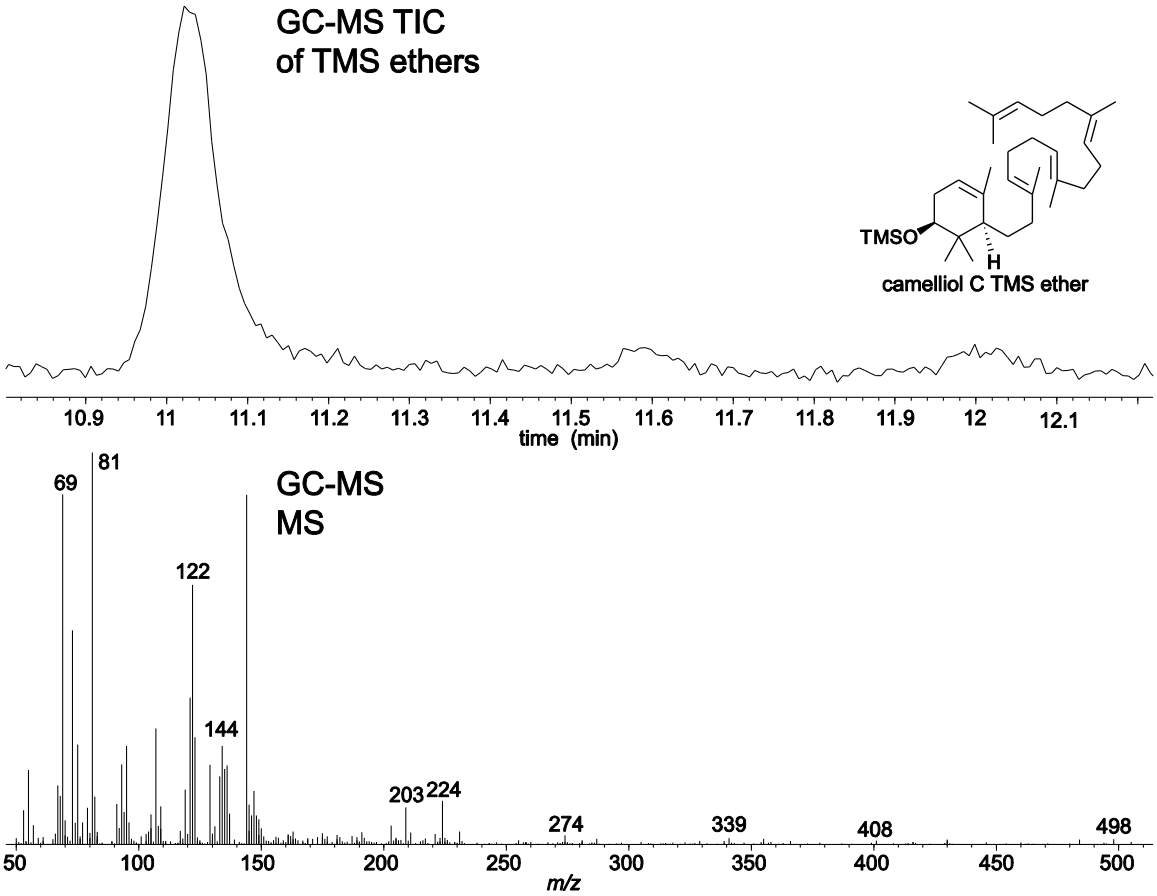


Figure F1.13. GC-MS TIC and EI mass spectra of camelliol C.

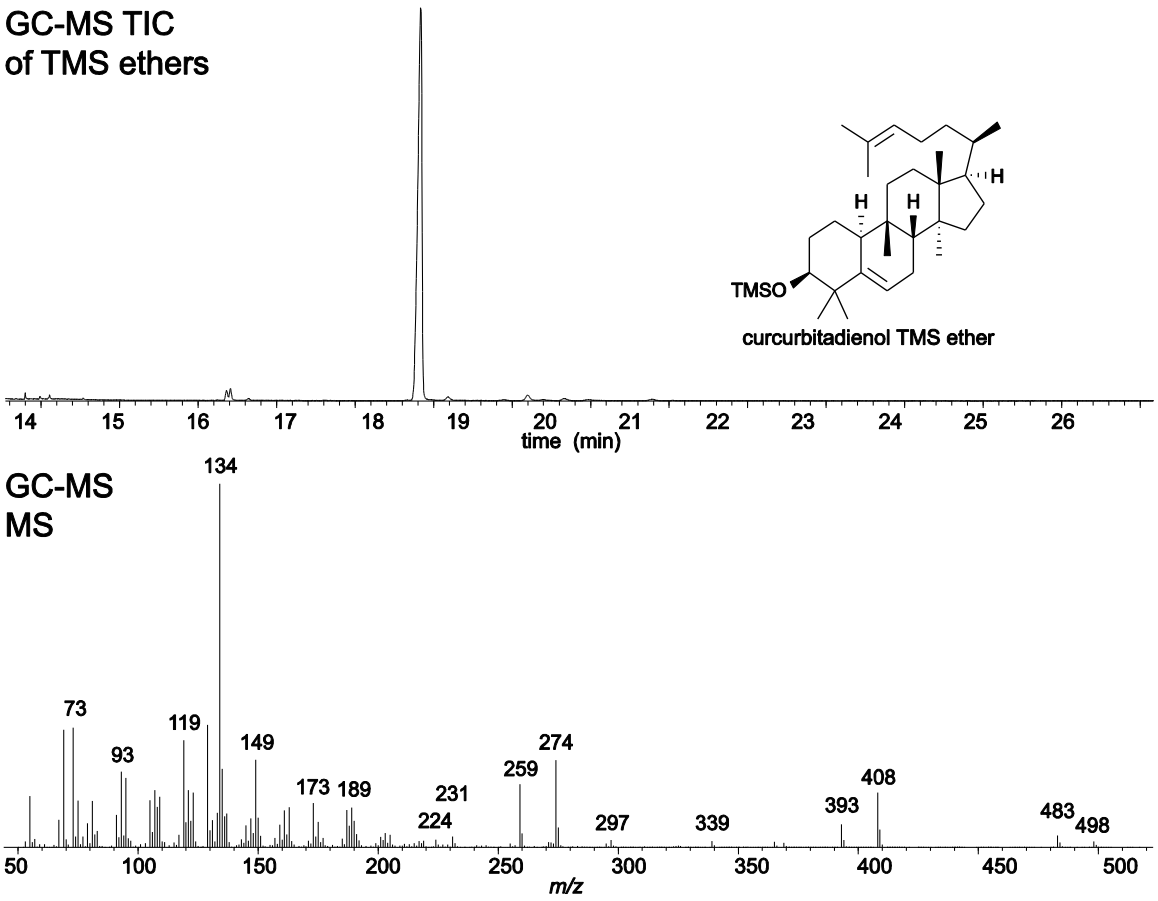


Figure F1.14. GC-MS TIC and EI mass spectra of curcurbitadienol.

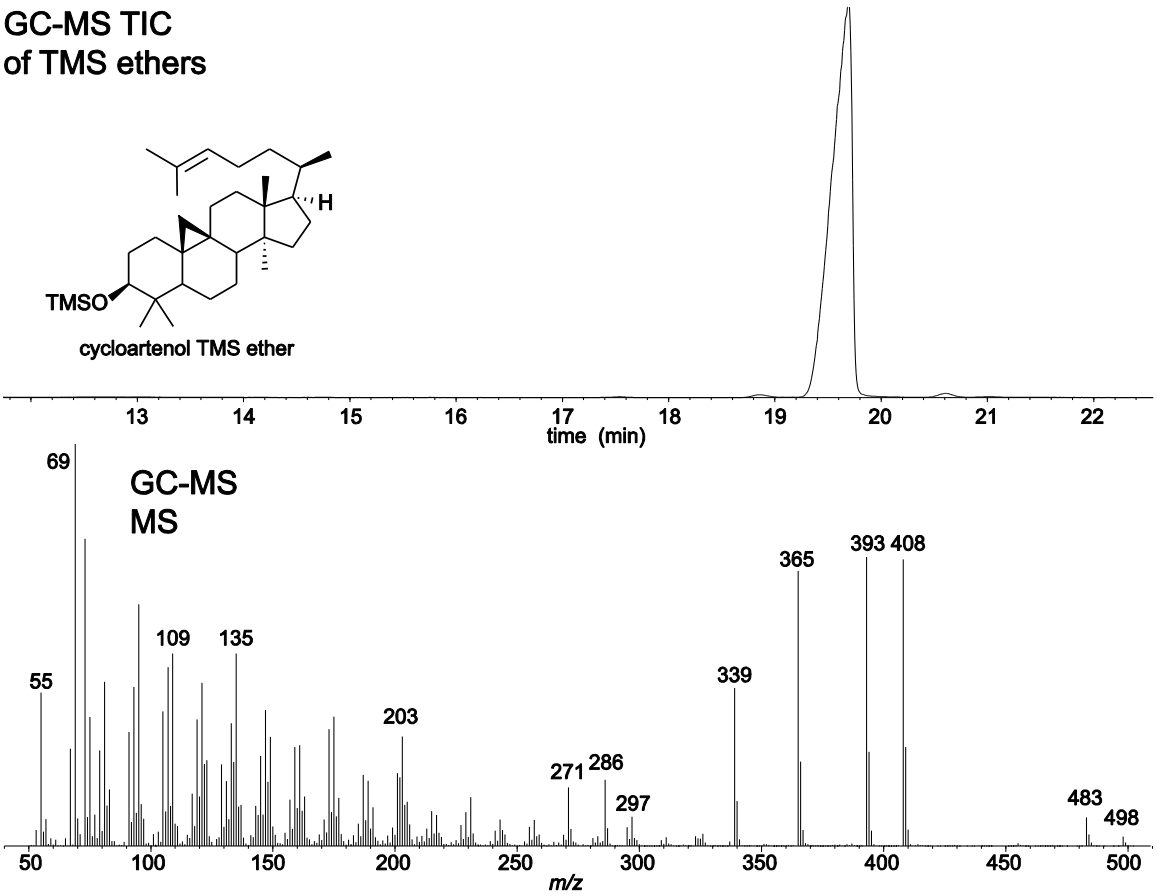


Figure F1.15. GC-MS TIC and EI mass spectra of cycloartenol.

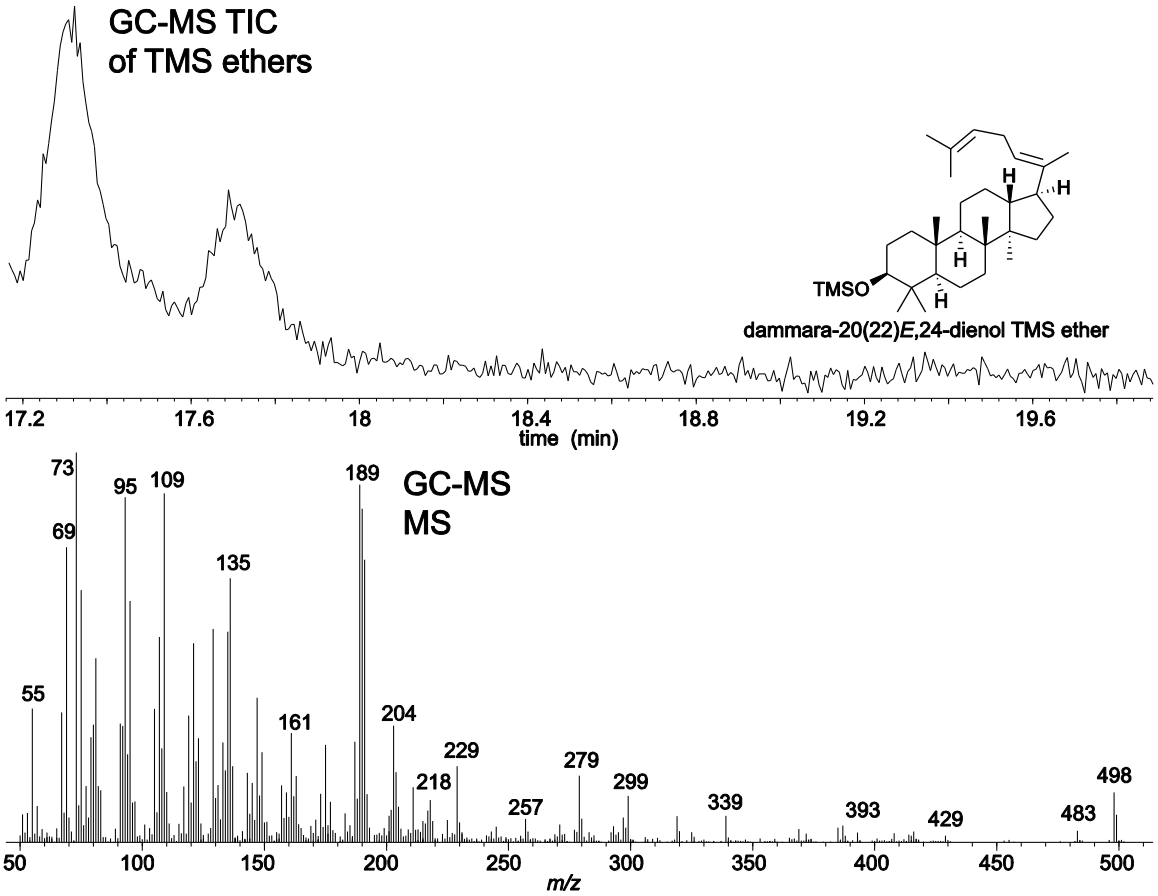


Figure F1.16. GC-MS TIC and EI mass spectra of dammara-20(22)*E*,24-dien-3β-ol.

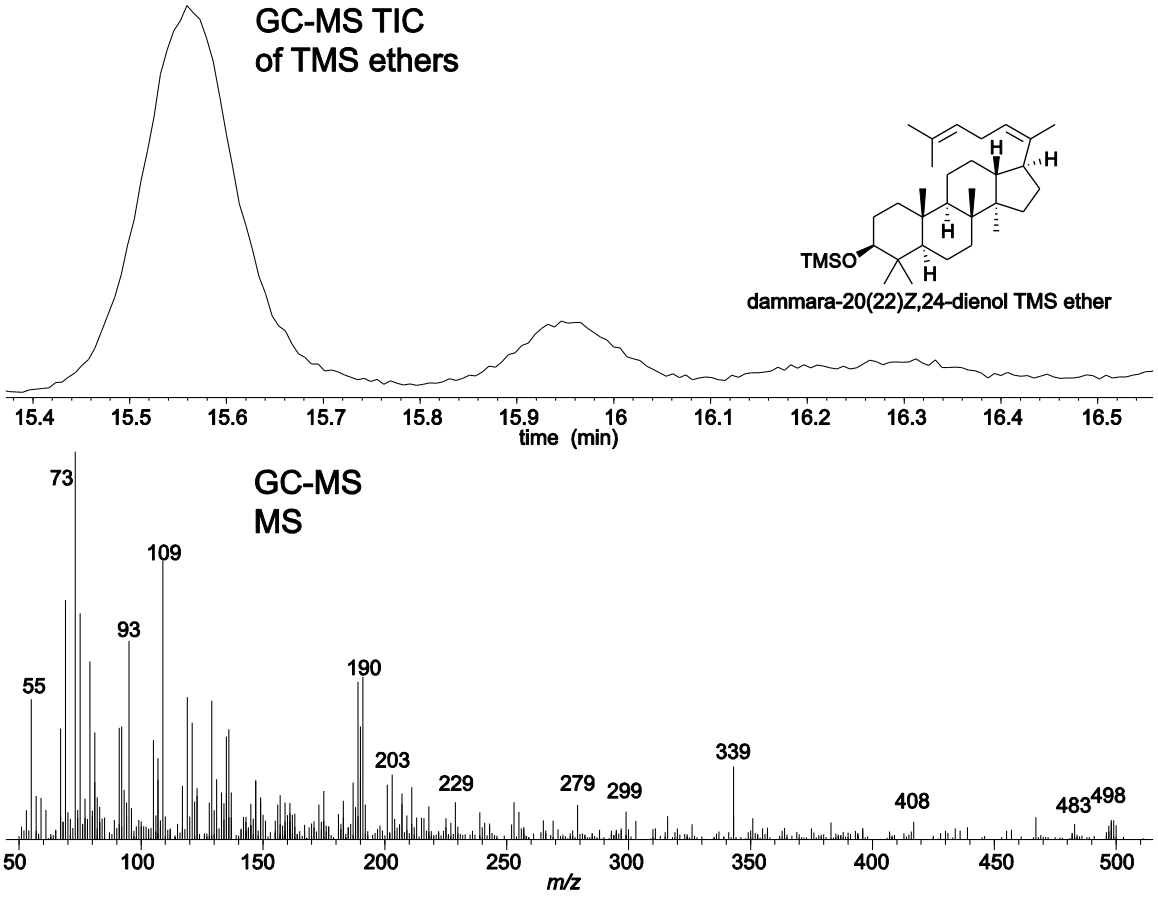


Figure F1.17. GC-MS TIC and EI mass spectra of dammara-20(22)*Z*,24-dien-3β-ol and dammara-20,24-dien-3β-ol.

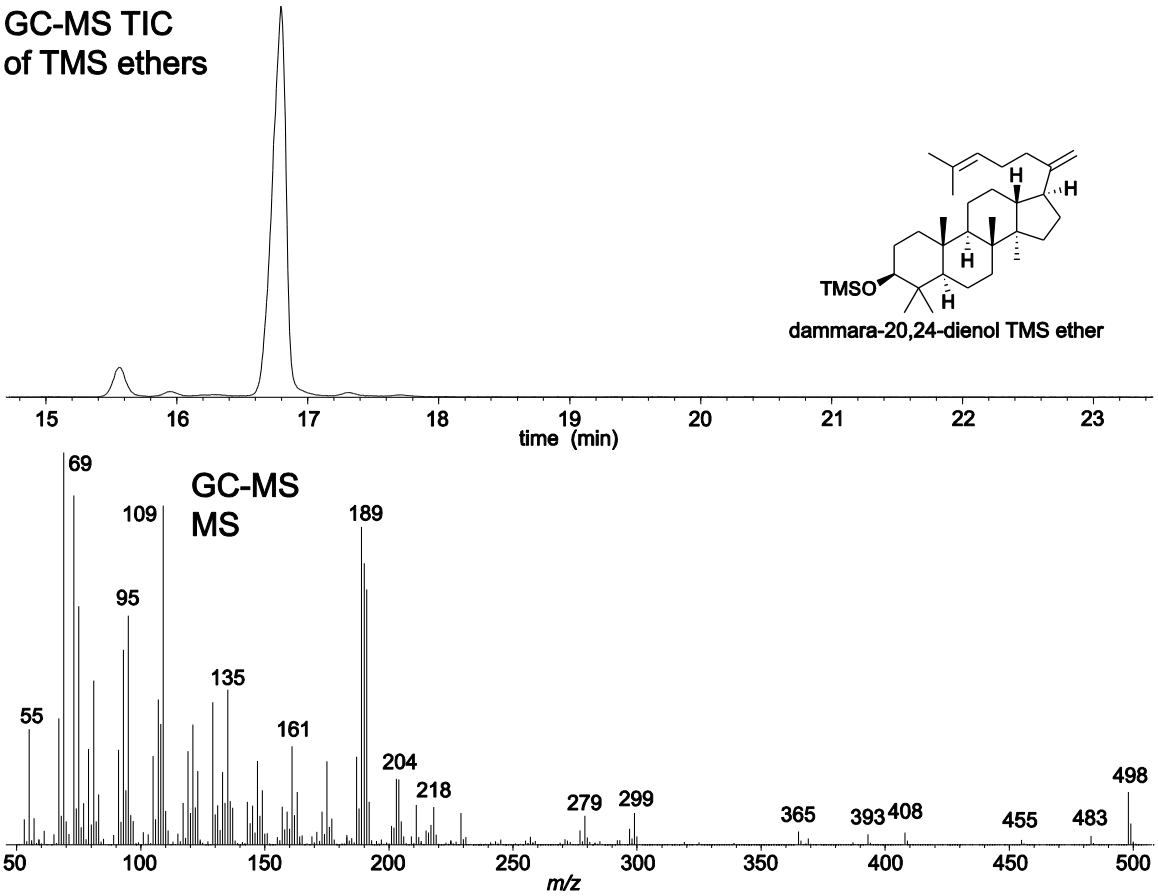


Figure F1.18. GC-MS TIC and EI mass spectra of dammara-20,24-dien-3β-ol.

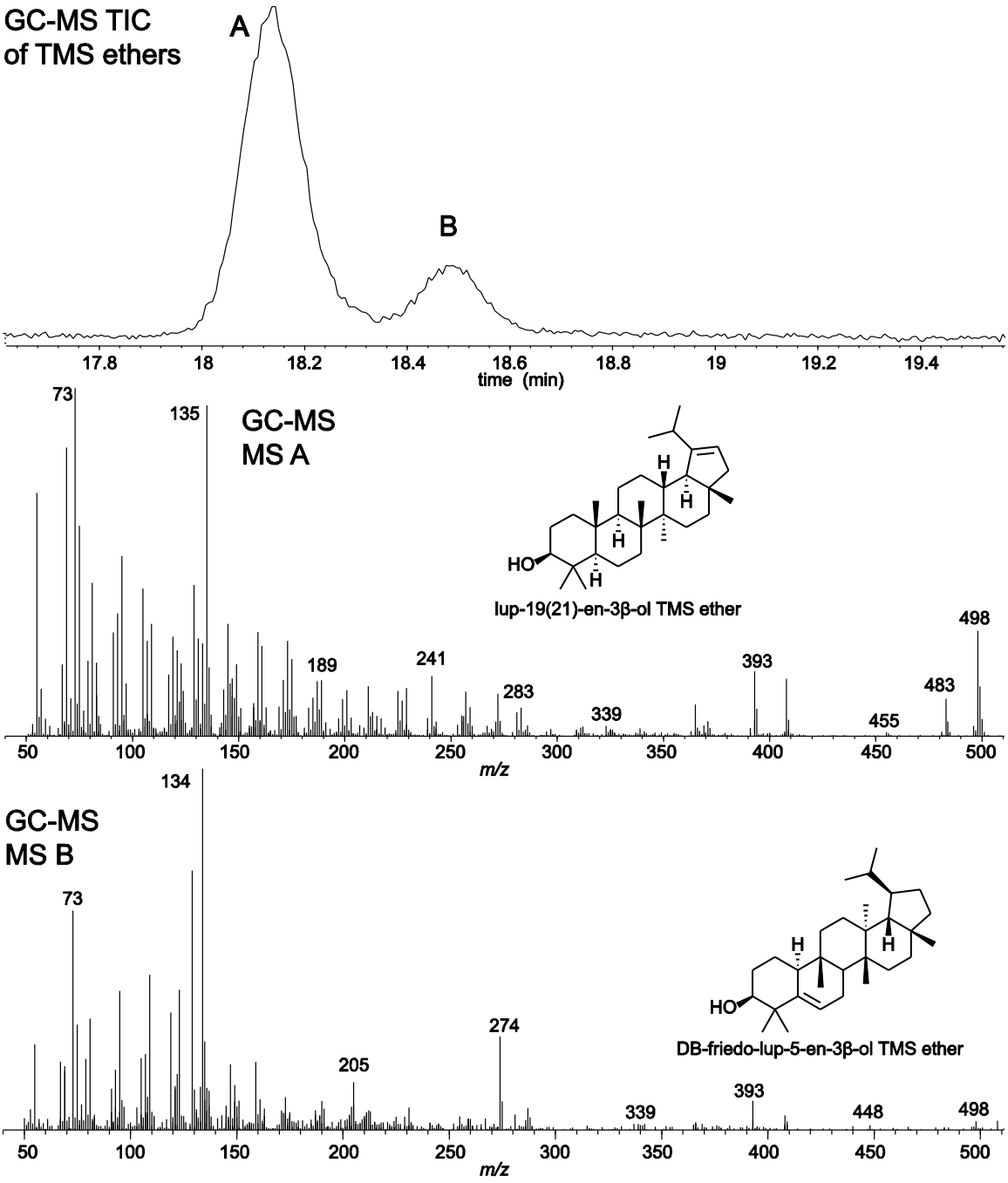


Figure F1.19. GC-MS TIC and EI mass spectra of DB-friedo-lup-5-en-3β-ol and lup-19(21)-en-3β-ol.

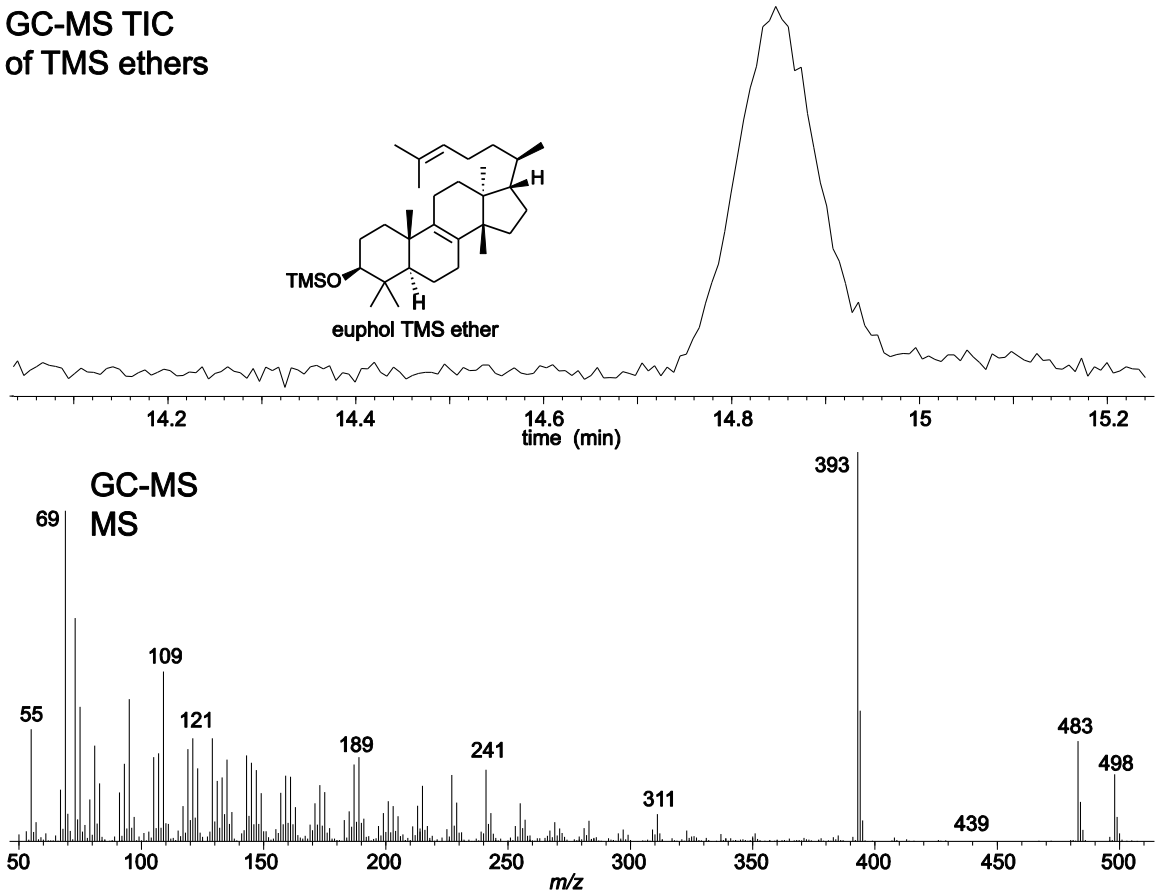


Figure F1.20. GC-MS TIC and EI mass spectra of euphol.

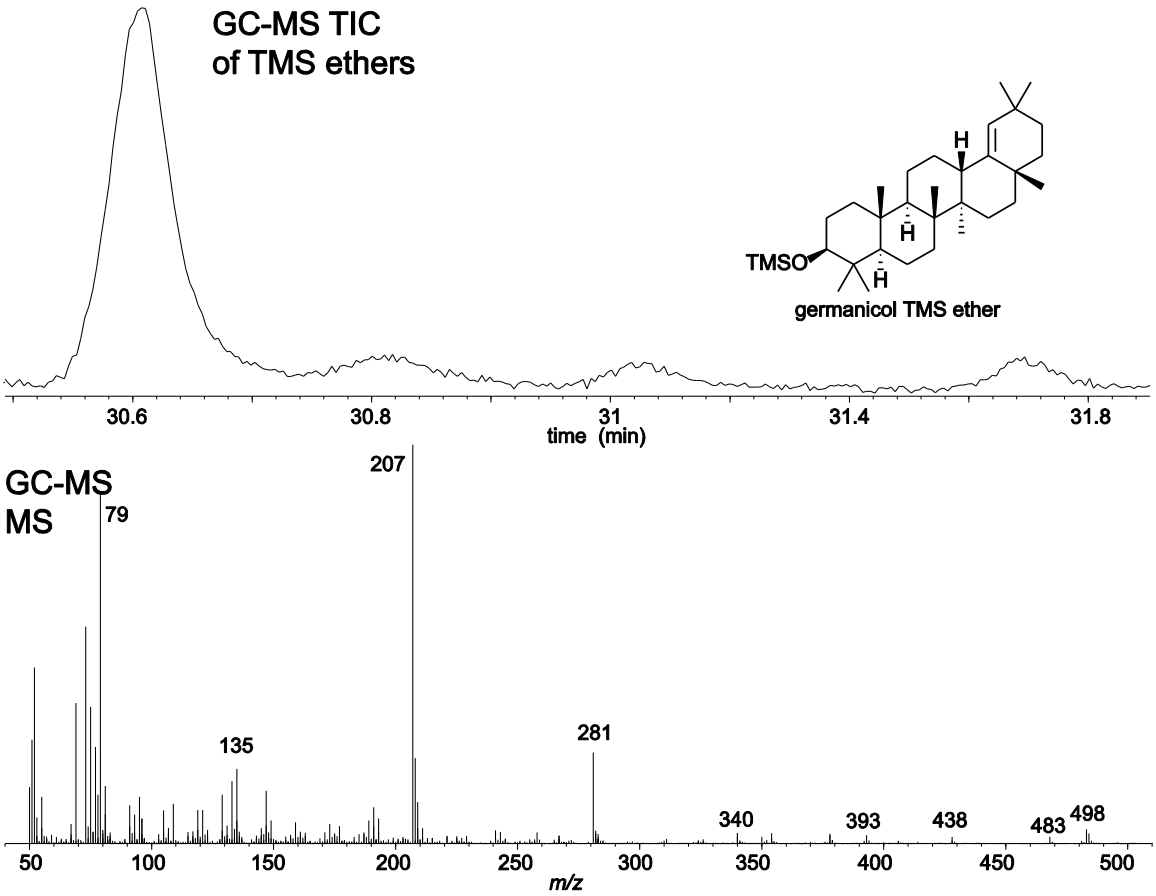


Figure F1.21. GC-MS TIC and EI mass spectra of germanicol.

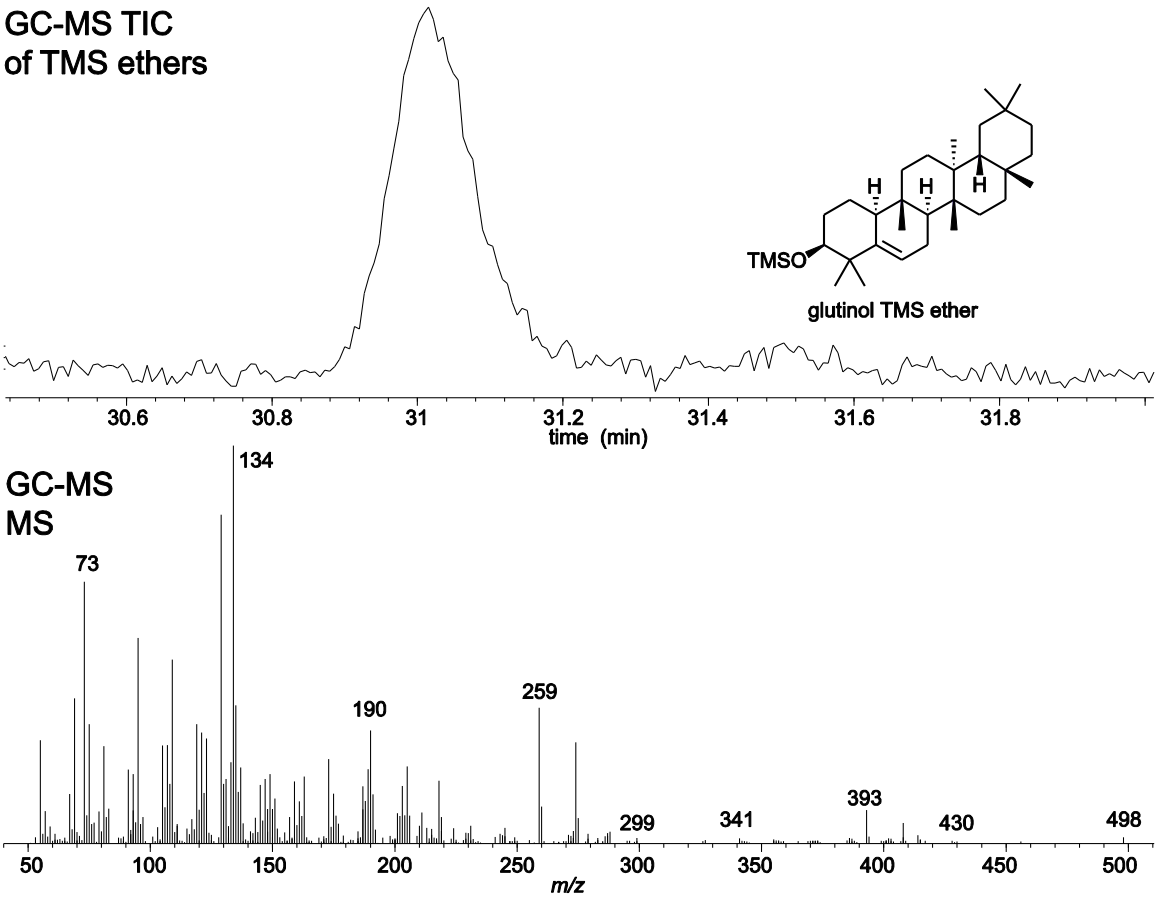


Figure F1.22. GC-MS TIC and EI mass spectra of glutinol.

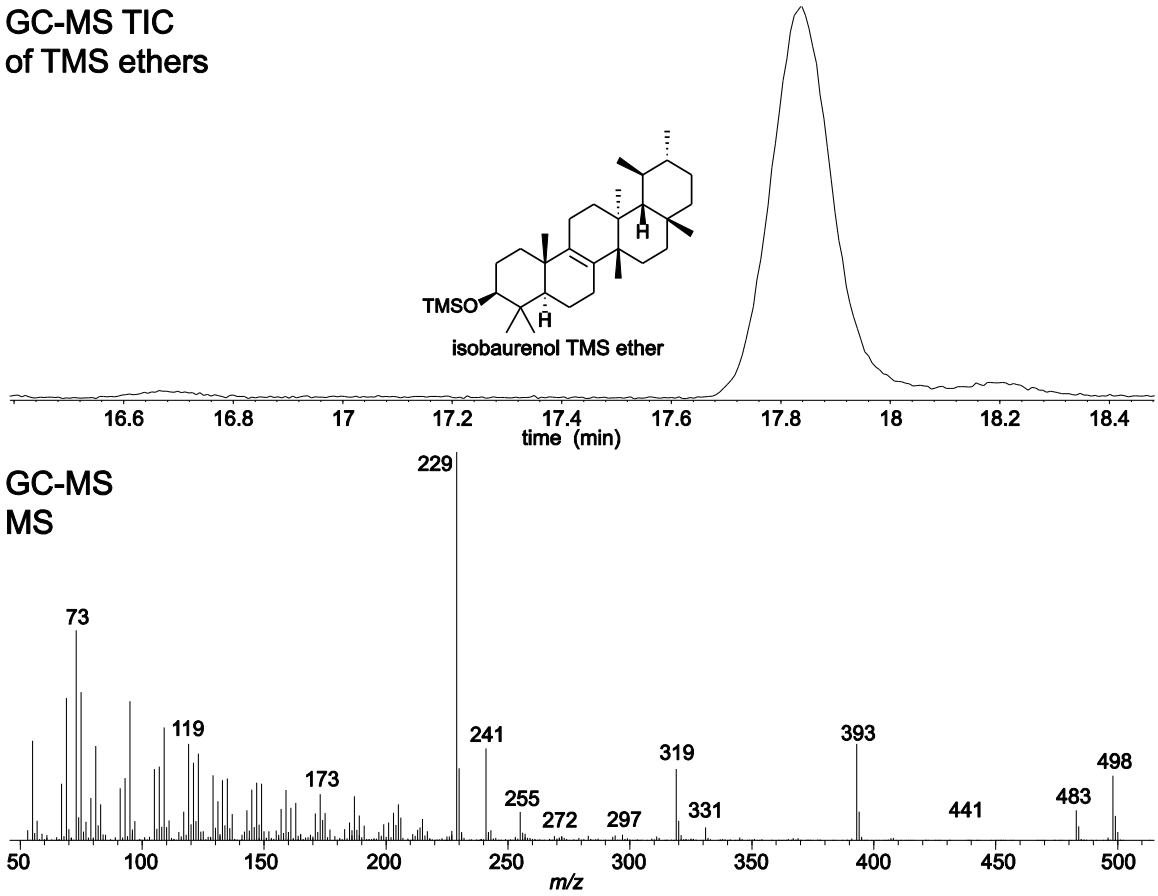


Figure F1.23. GC-MS TIC and EI mass spectra of isobauerenol.

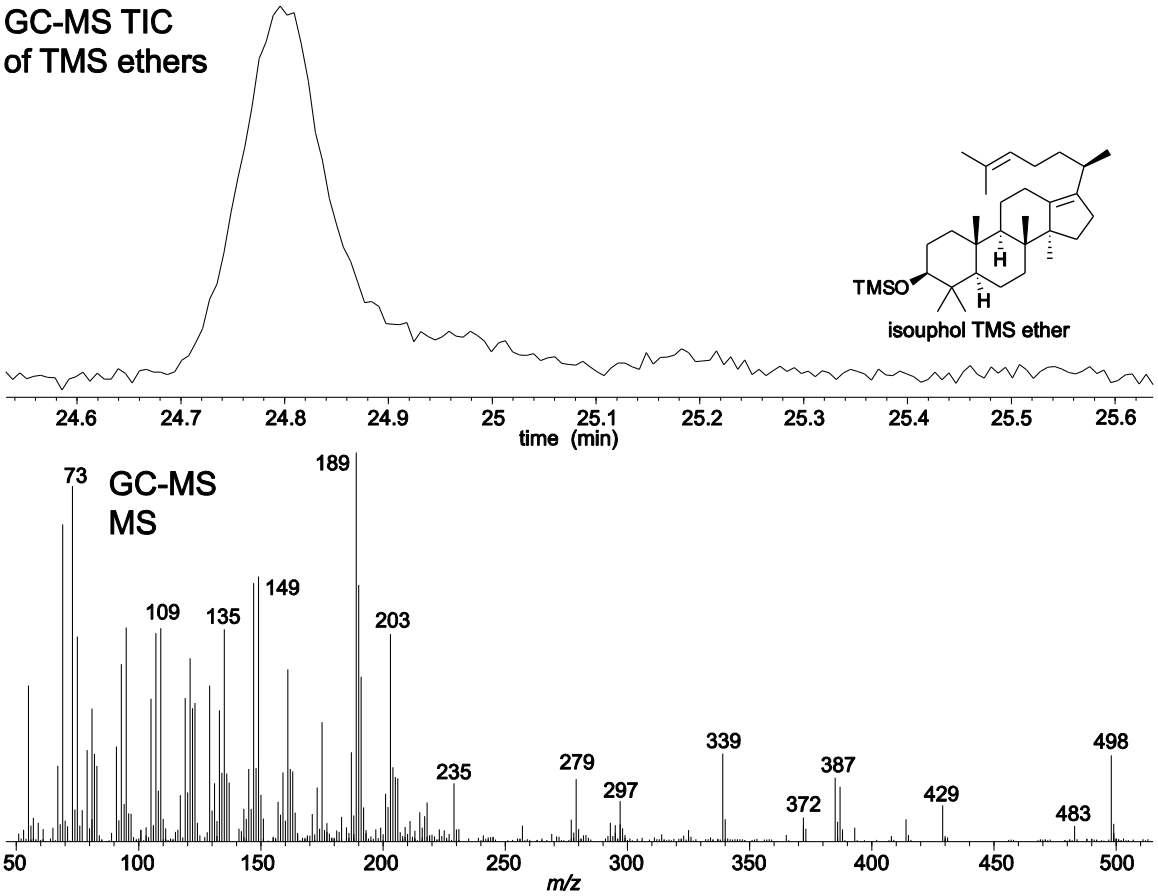


Figure F1.24. GC-MS TIC and EI mass spectra of isoeuphol.

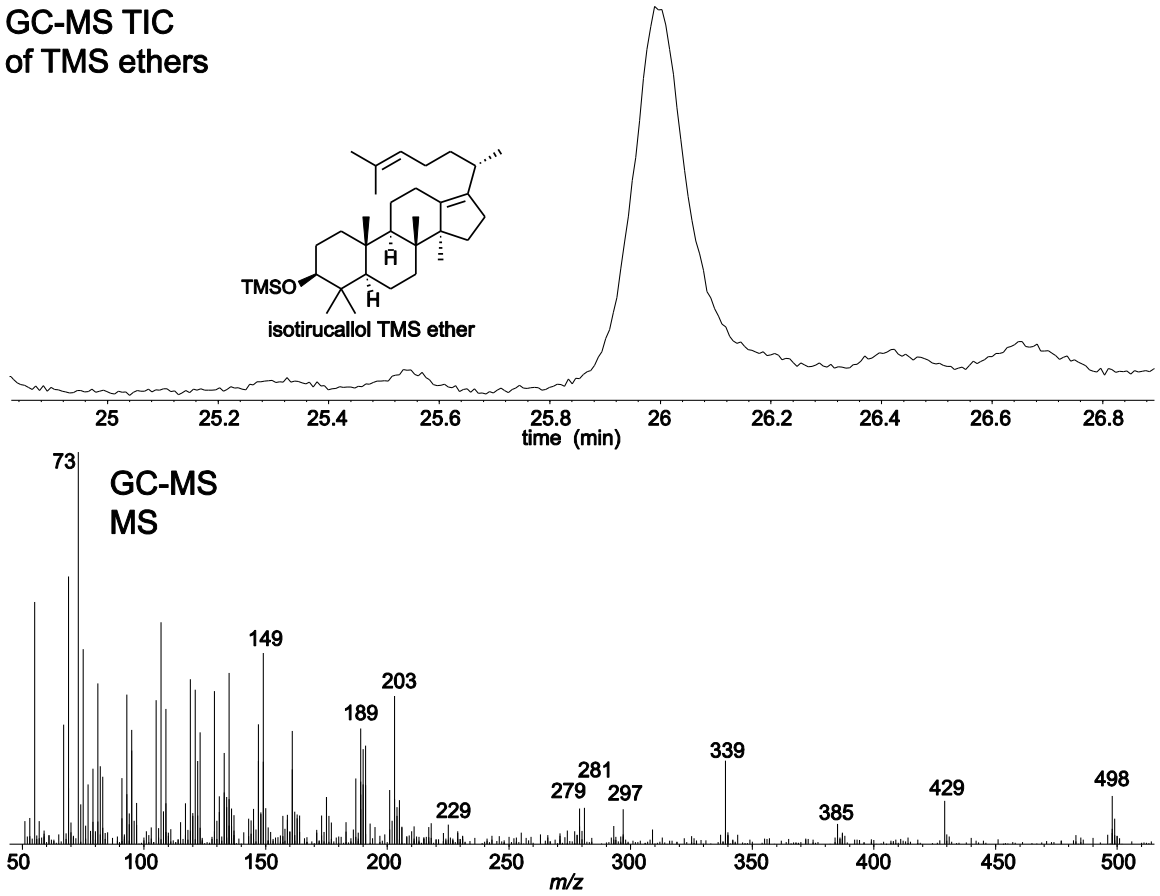


Figure F1.25. GC-MS TIC and EI mass spectra of isotirucallol.

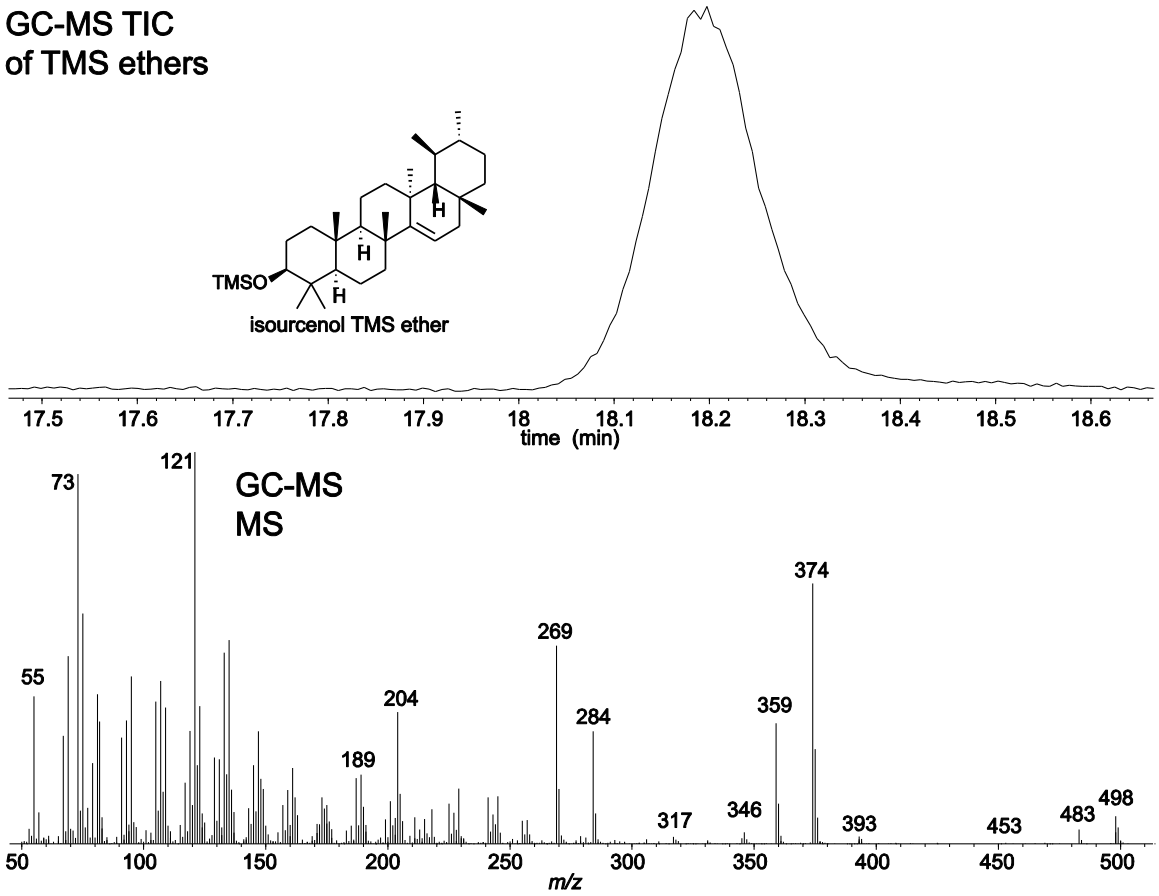


Figure F1.26. GC-MS TIC and EI mass spectra of isoursenol.

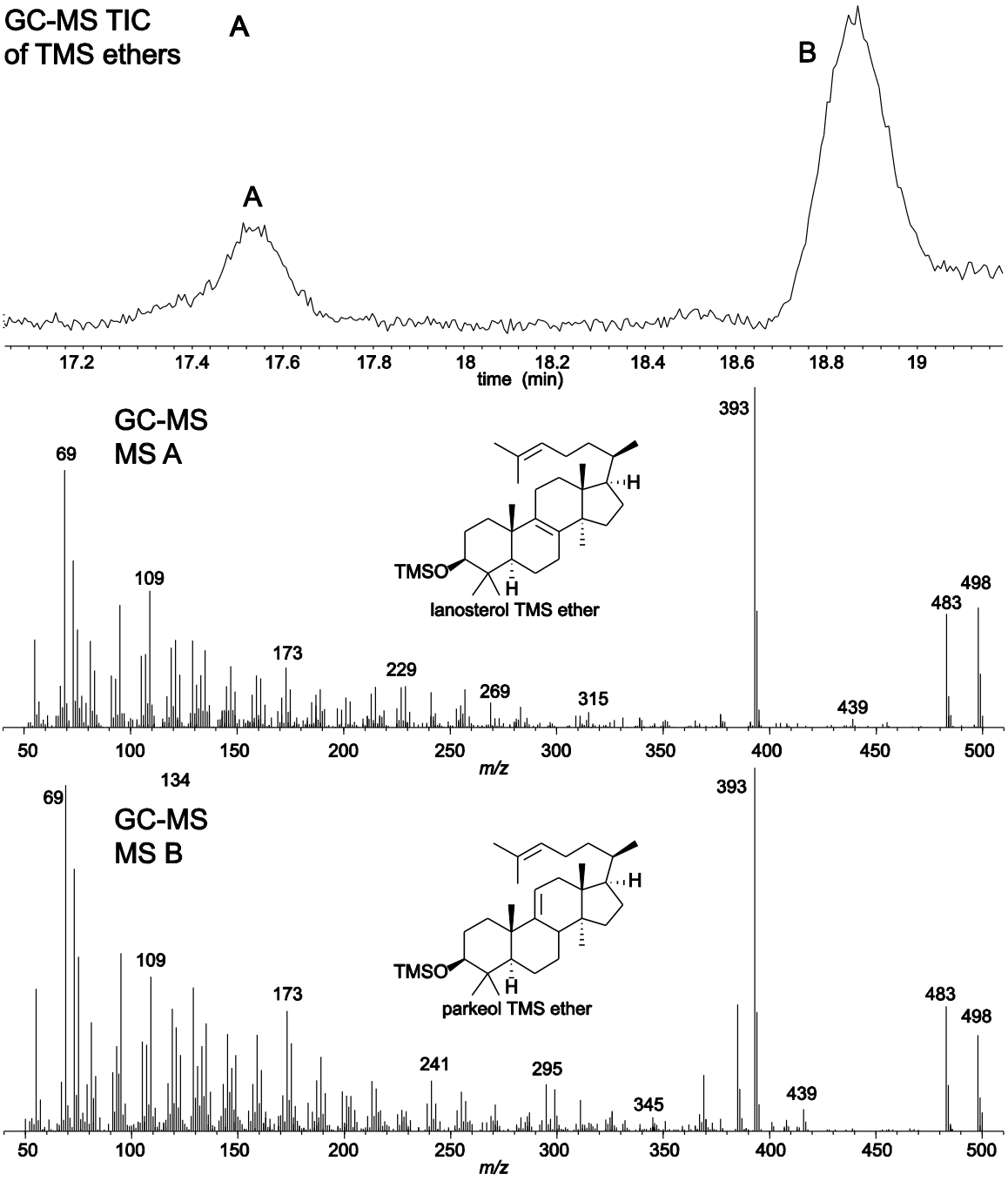


Figure F1.27. GC-MS TIC and EI mass spectra of lanosterol and parkeol.

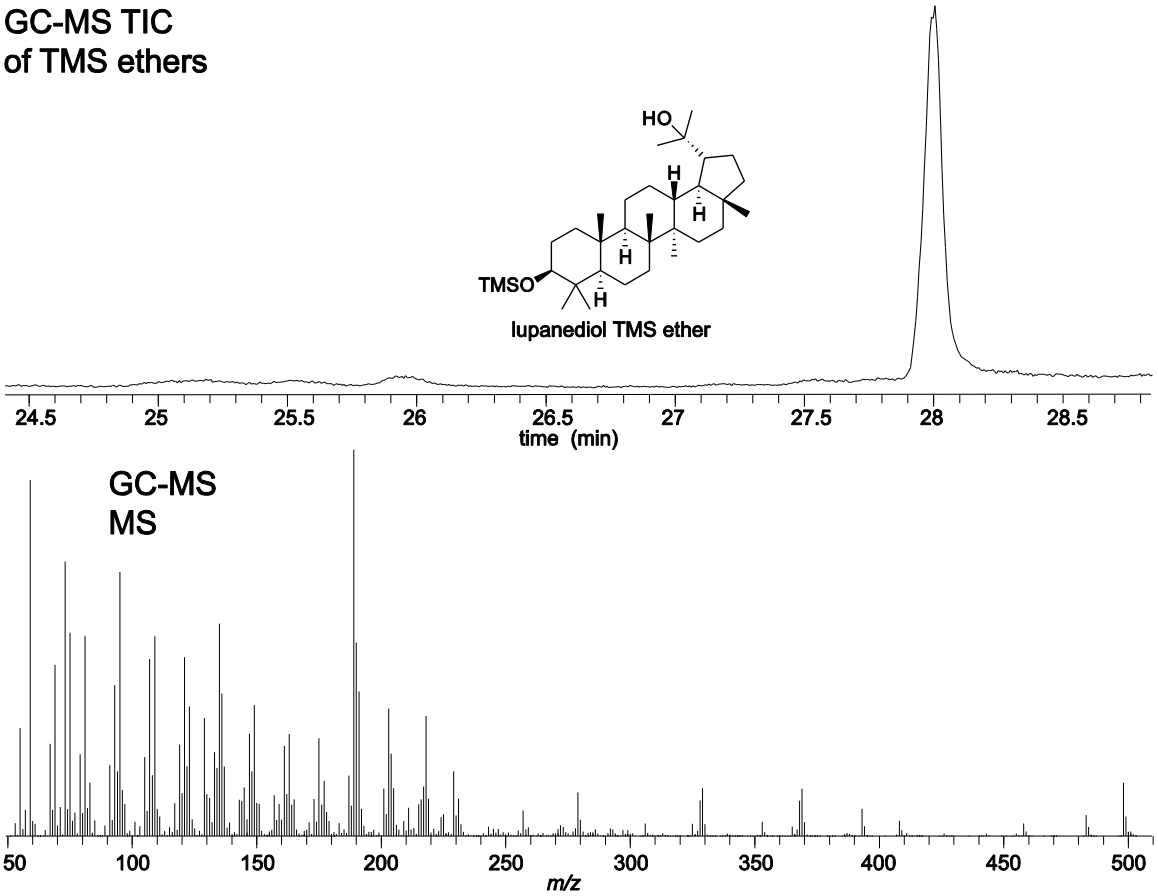


Figure F1.28. GC-MS TIC and EI mass spectra of lupane-3β,20-diol.

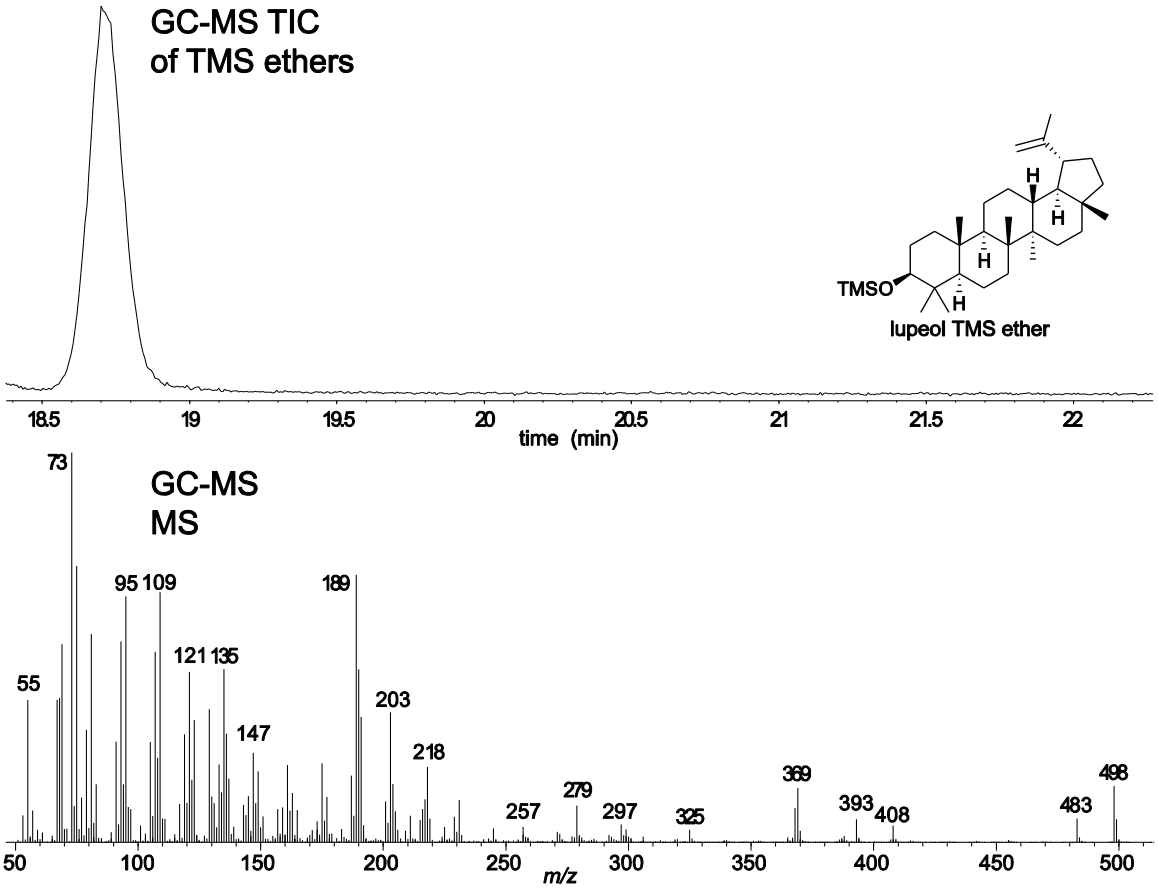


Figure F1.29. GC-MS TIC and EI mass spectra of lupeol.

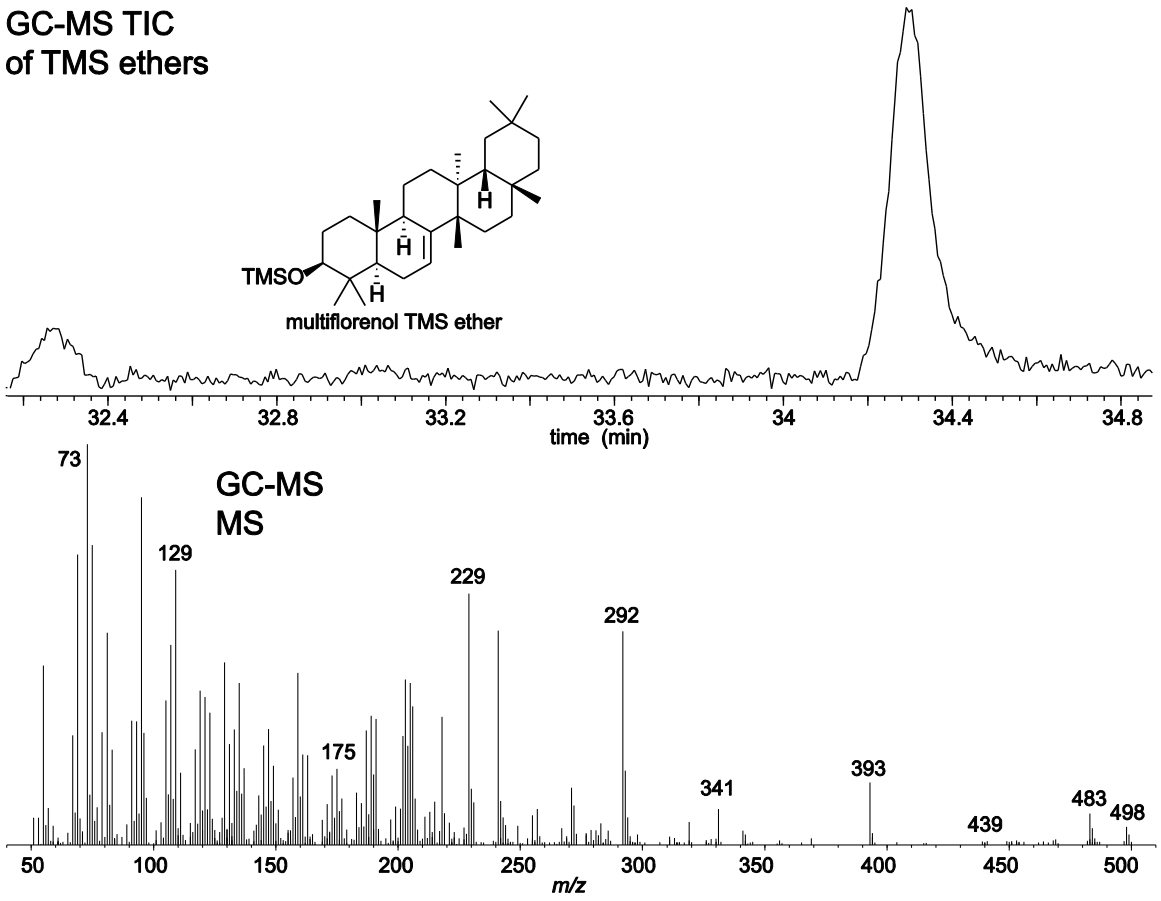


Figure F1.30. GC-MS TIC and EI mass spectra of multiflorenol.

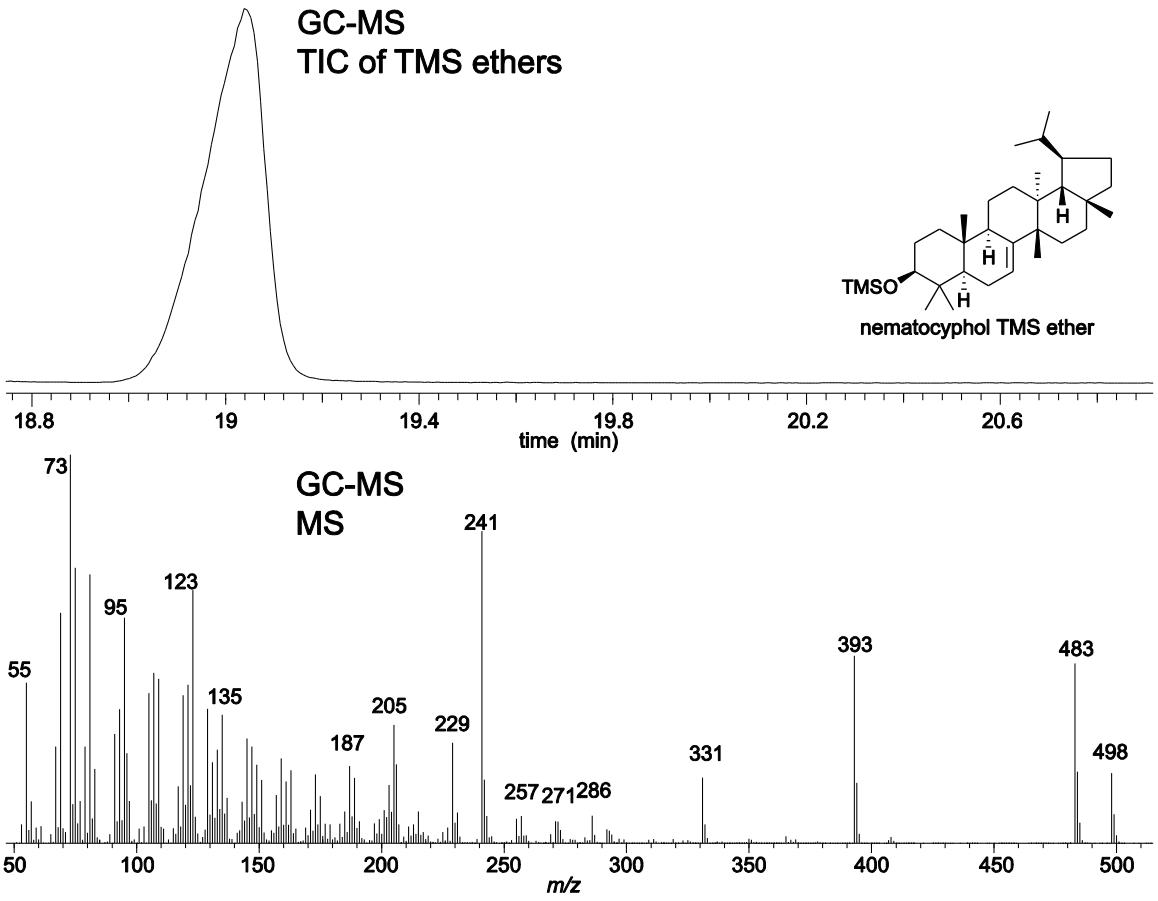


Figure F1.31. GC-MS TIC and EI mass spectra of nematocyphol.

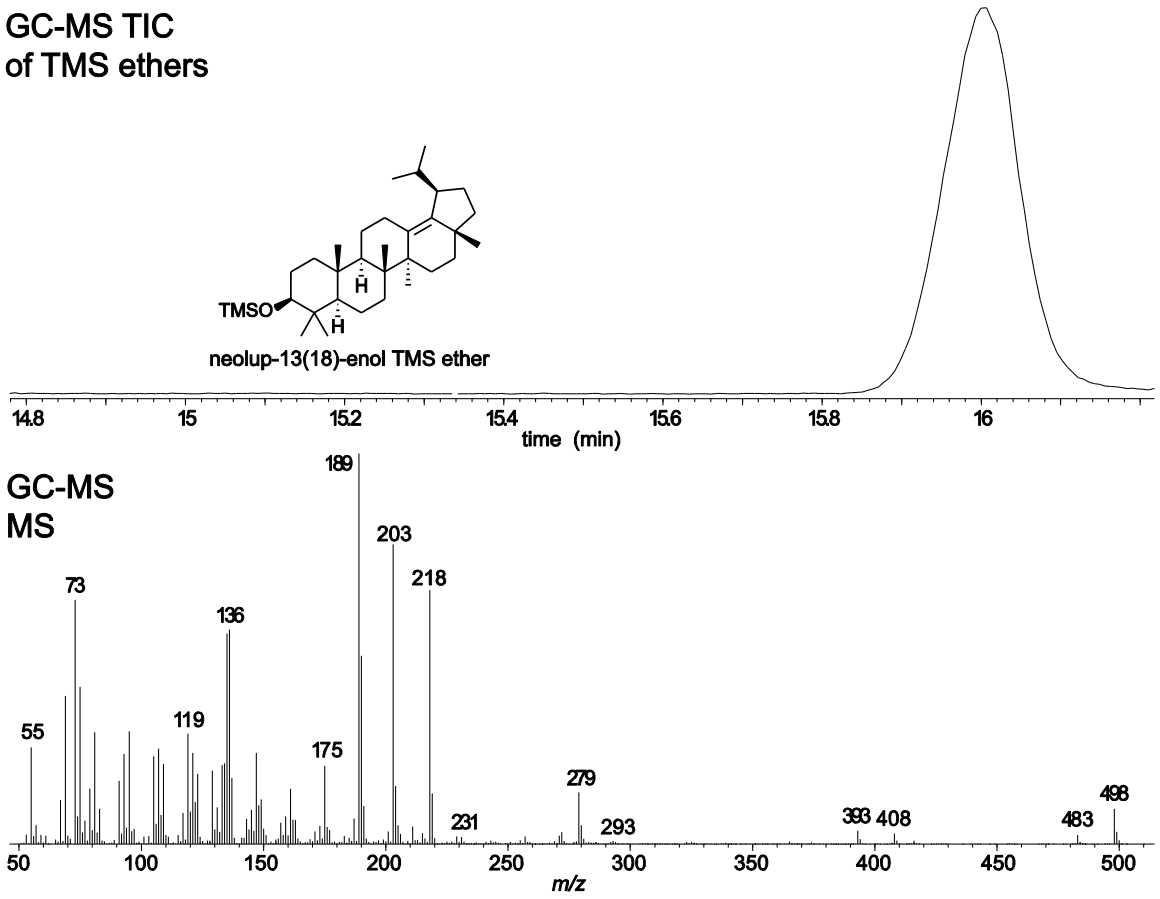


Figure F1.32. GC-MS TIC and EI mass spectra of neolup-13(18)-enol.

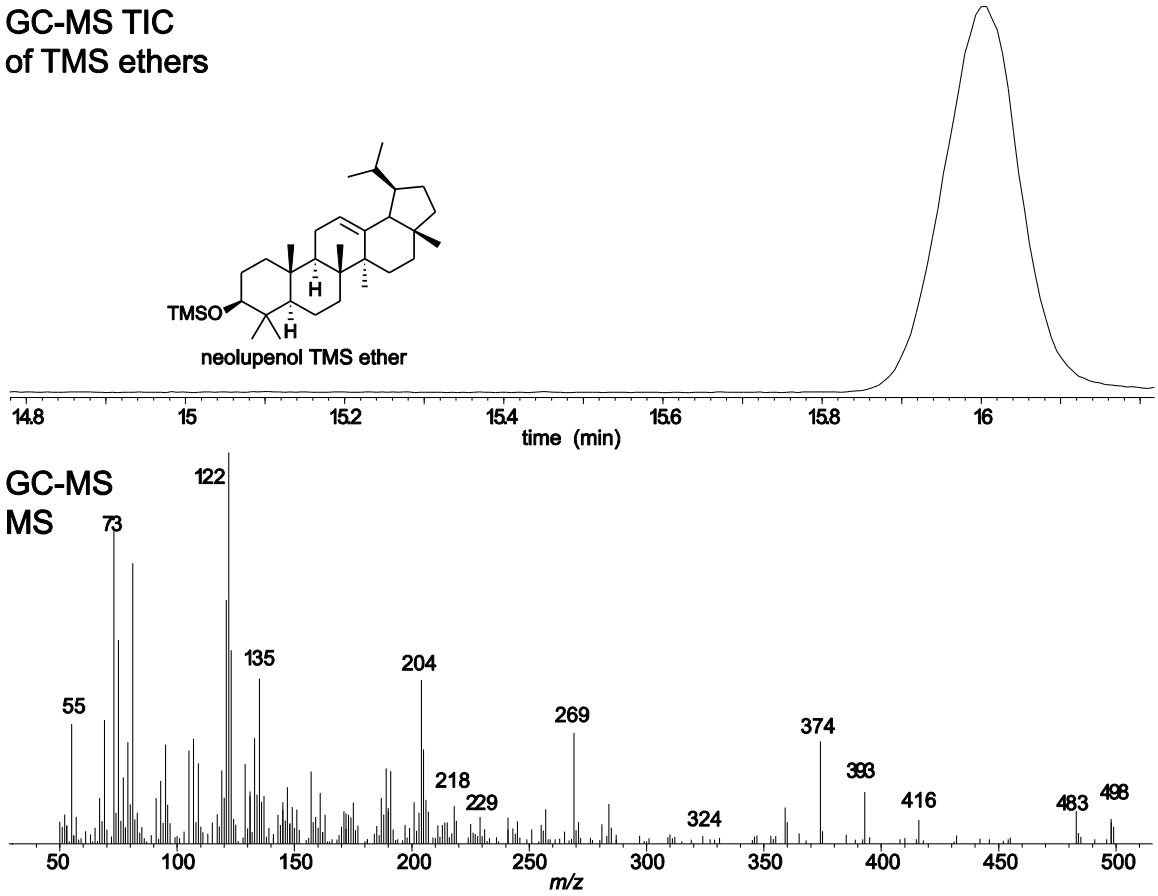


Figure F1.33. GC-MS TIC and EI mass spectra of neolupenol.

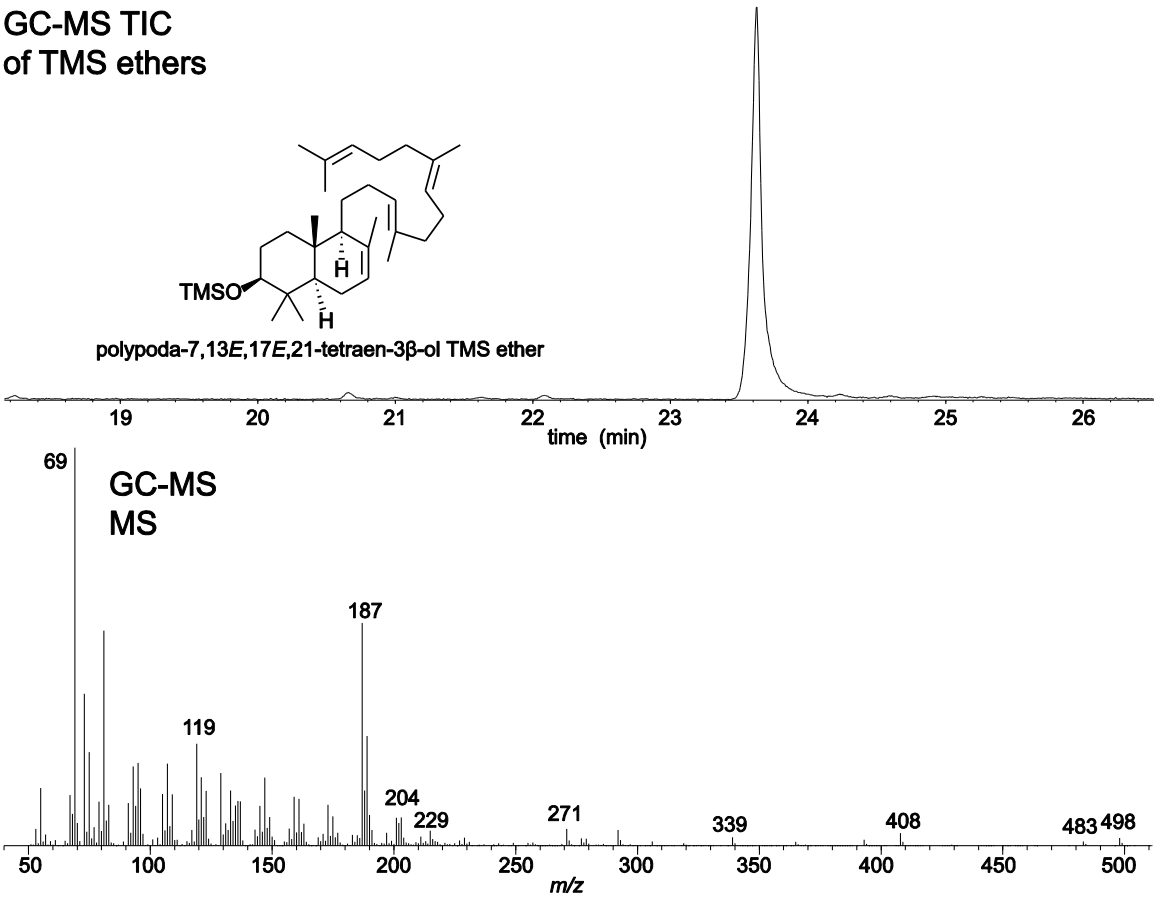


Figure F1.34. GC-MS TIC and EI mass spectra of polypoda-7,13*E*,17*E*,21-tetraen-3β-ol.

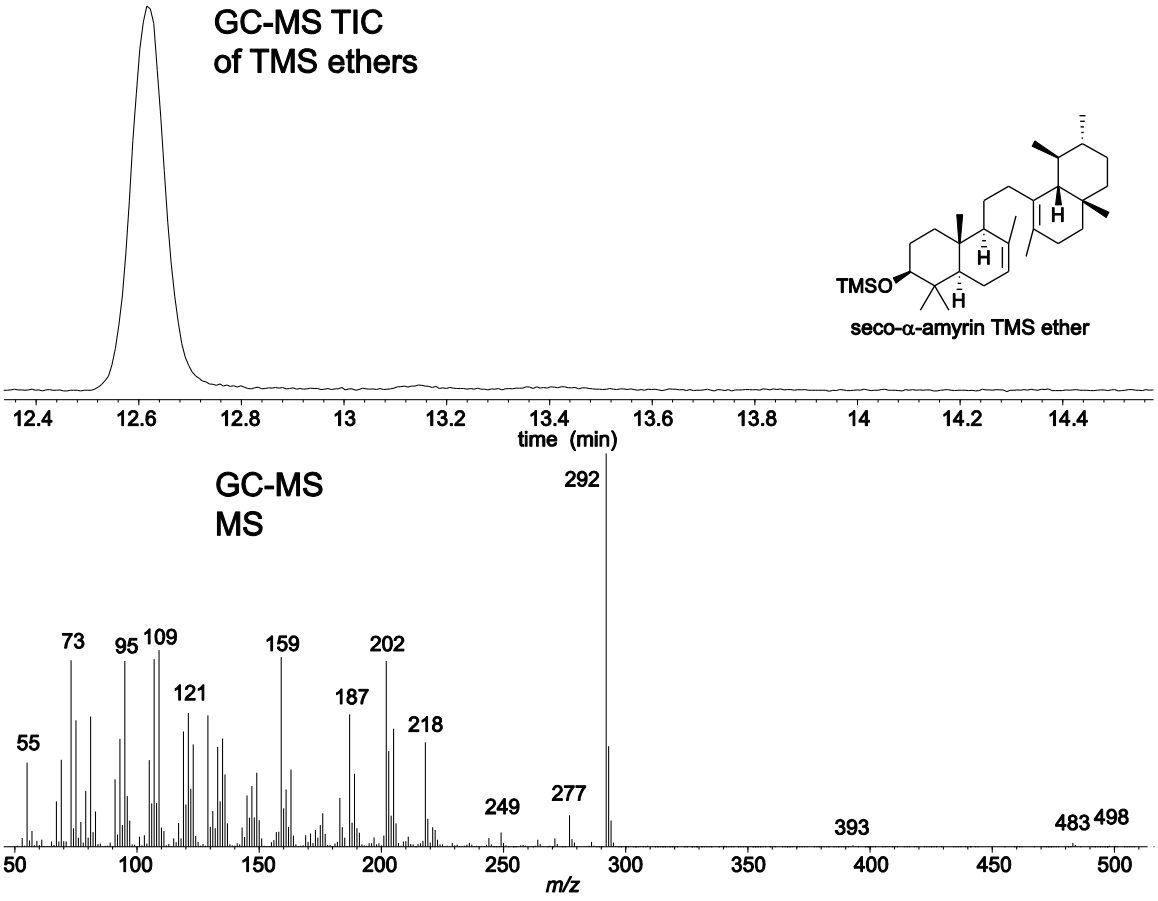


Figure F1.35. GC-MS TIC and EI mass spectra of seco-α-amyrin.

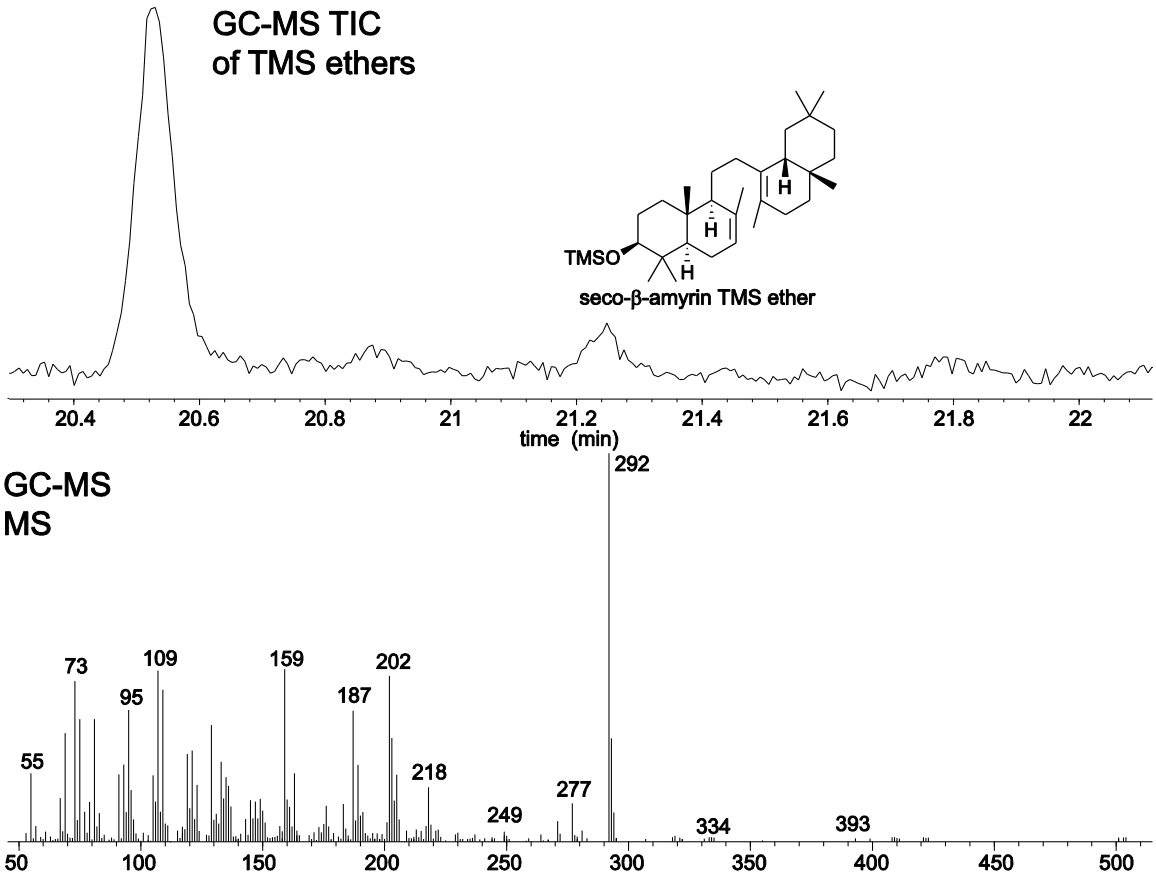


Figure F1.36. GC-MS TIC and EI mass spectra of seco-β-amyrin.

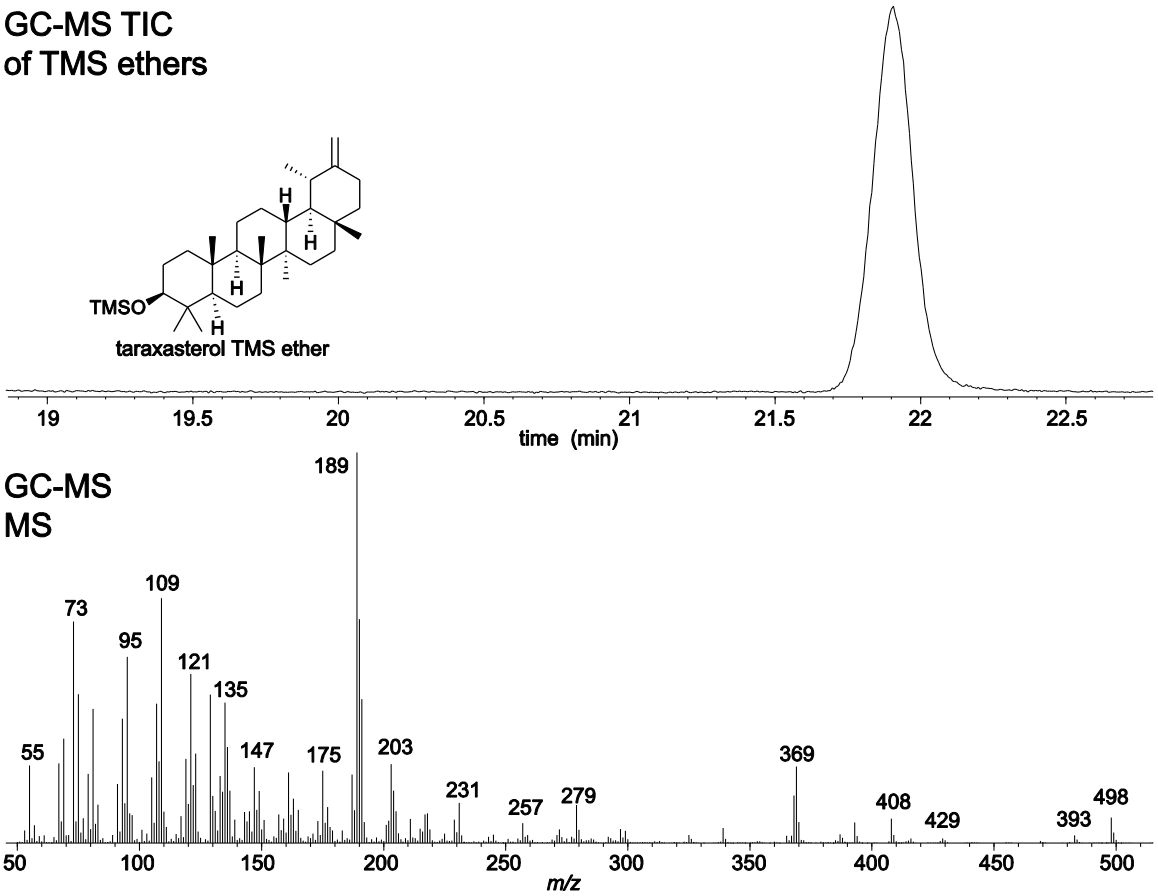


Figure F1.37. GC-MS TIC and EI mass spectra of taraxasterol.

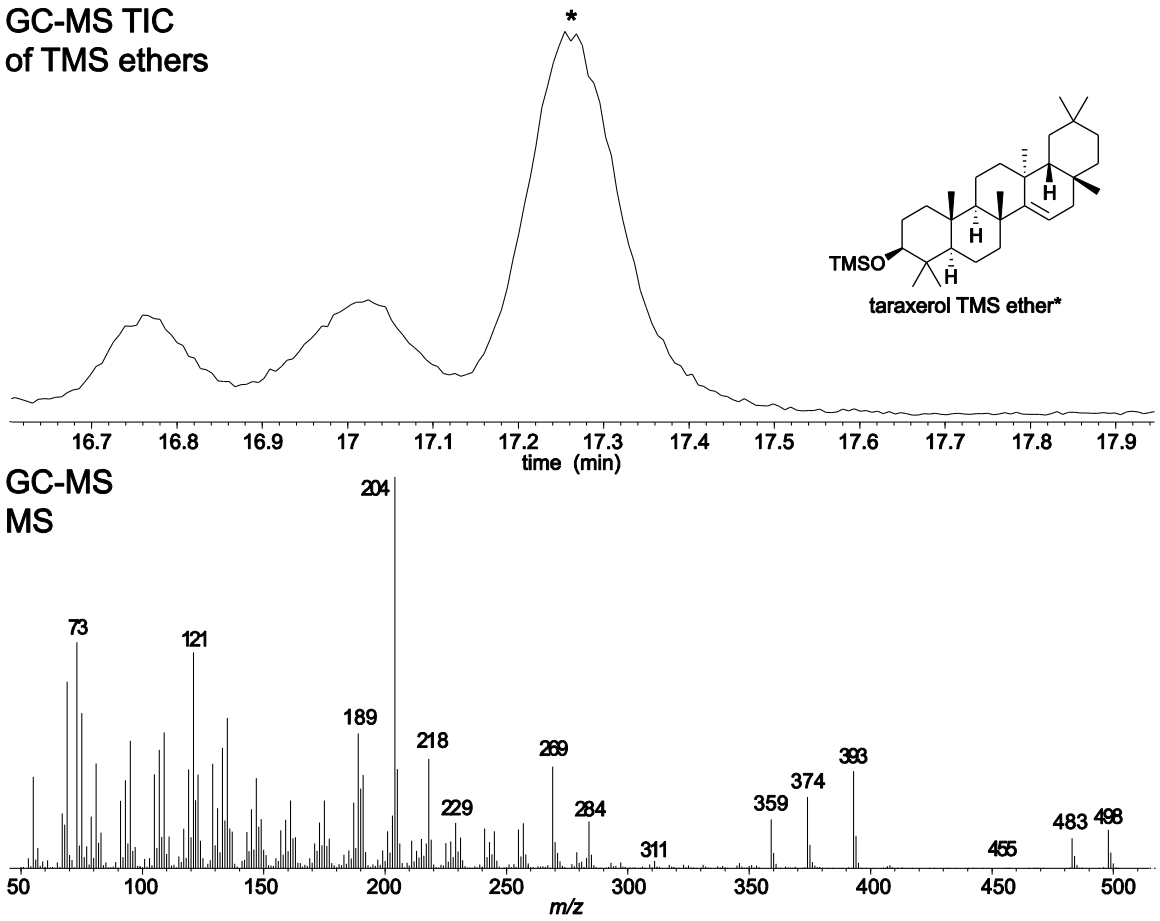


Figure F1.38. GC-MS TIC and EI mass spectra of taraxerol.

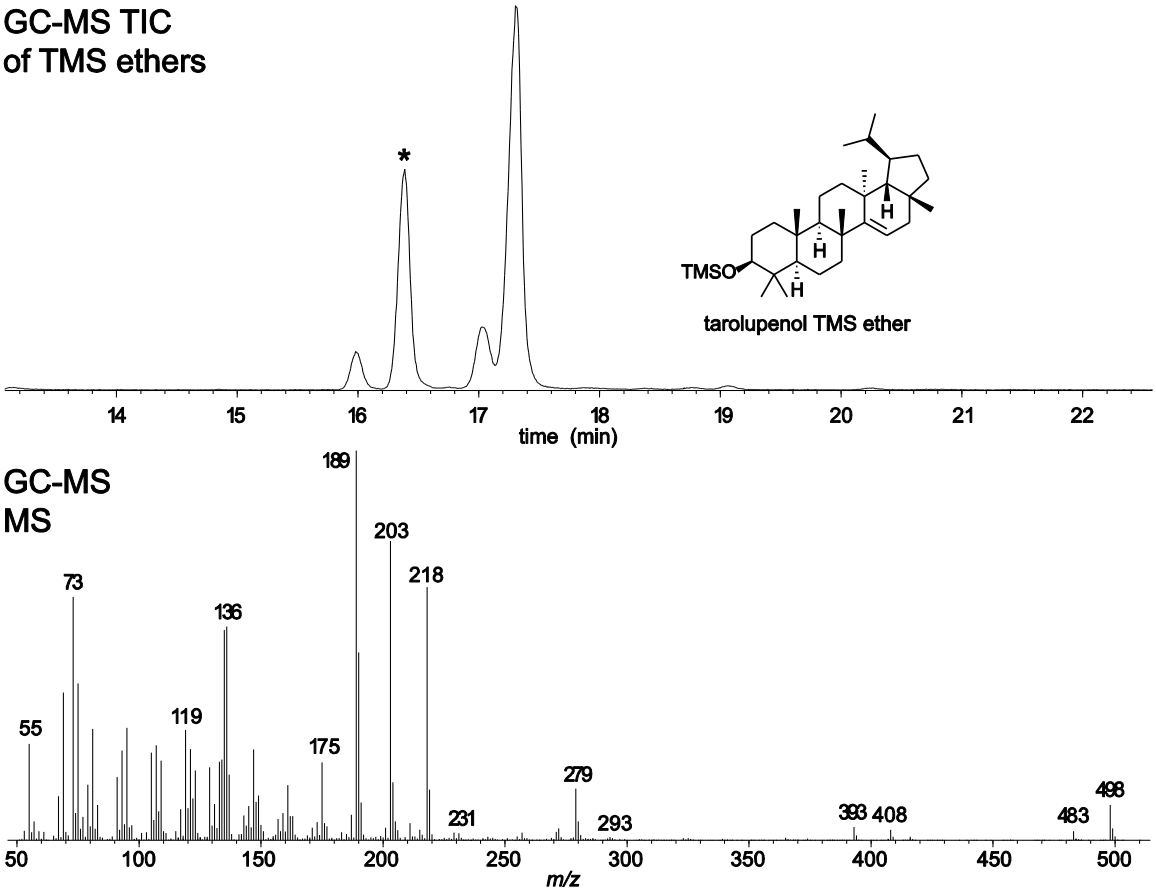


Figure F1.39. GC-MS TIC and EI mass spectra of tarolupenol.

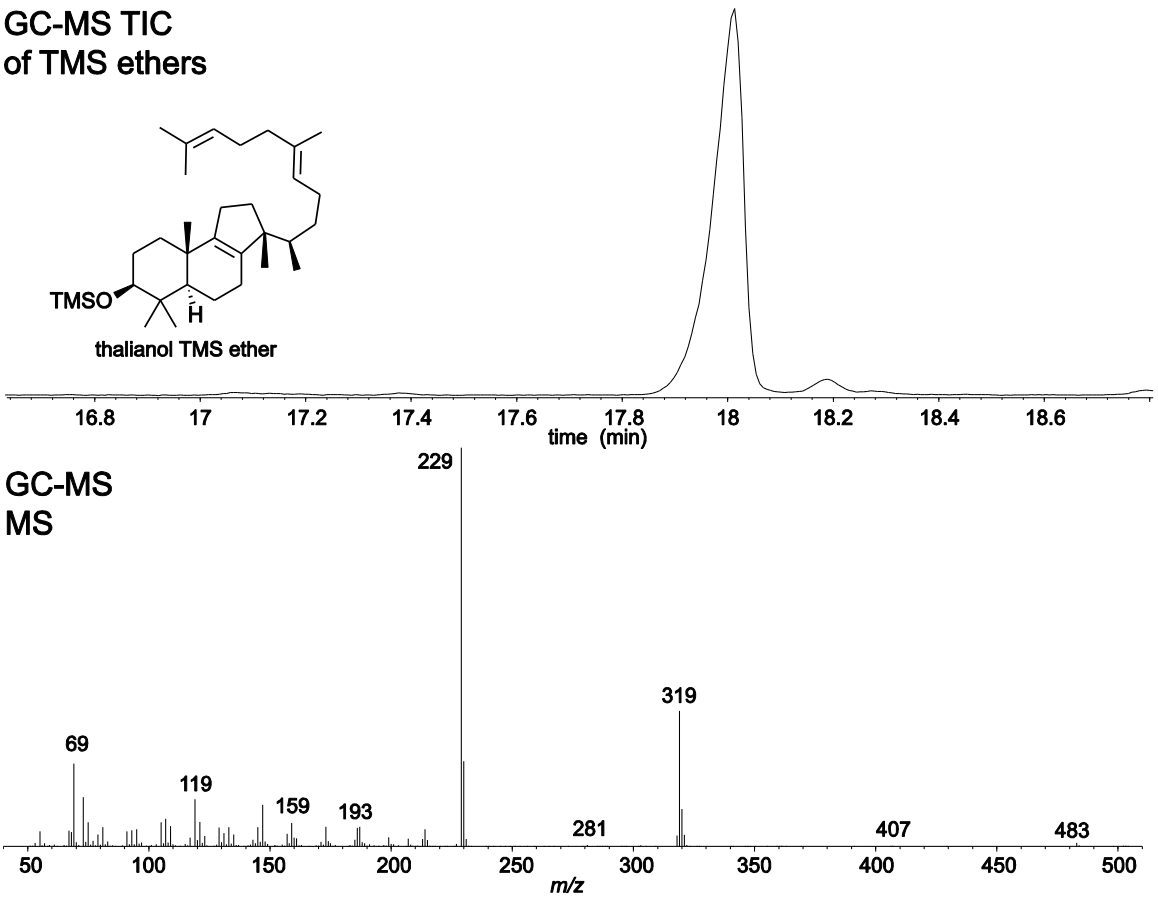


Figure F1.40. GC-MS TIC and EI mass spectra of thalianol.

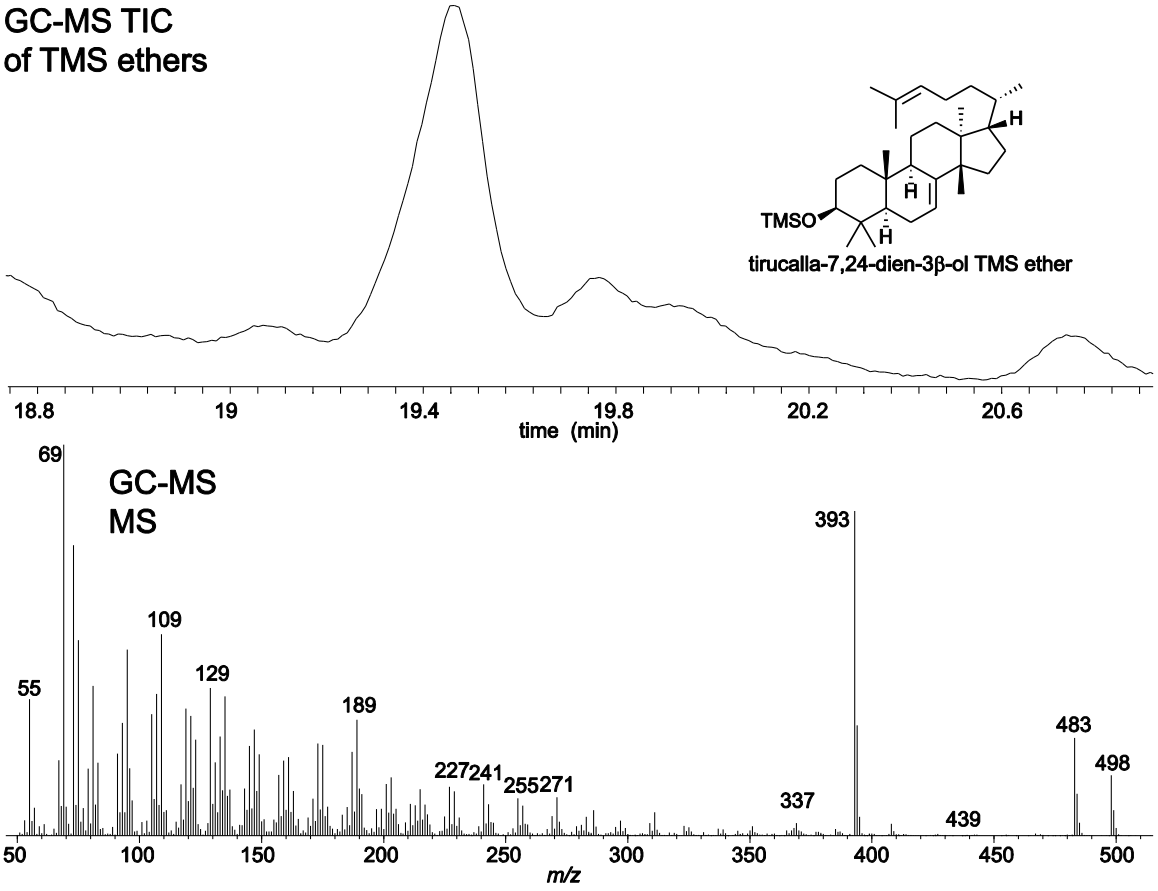


Figure F1.41. GC-MS TIC and EI mass spectra of tirucalla-7,24-dien-3β-ol.

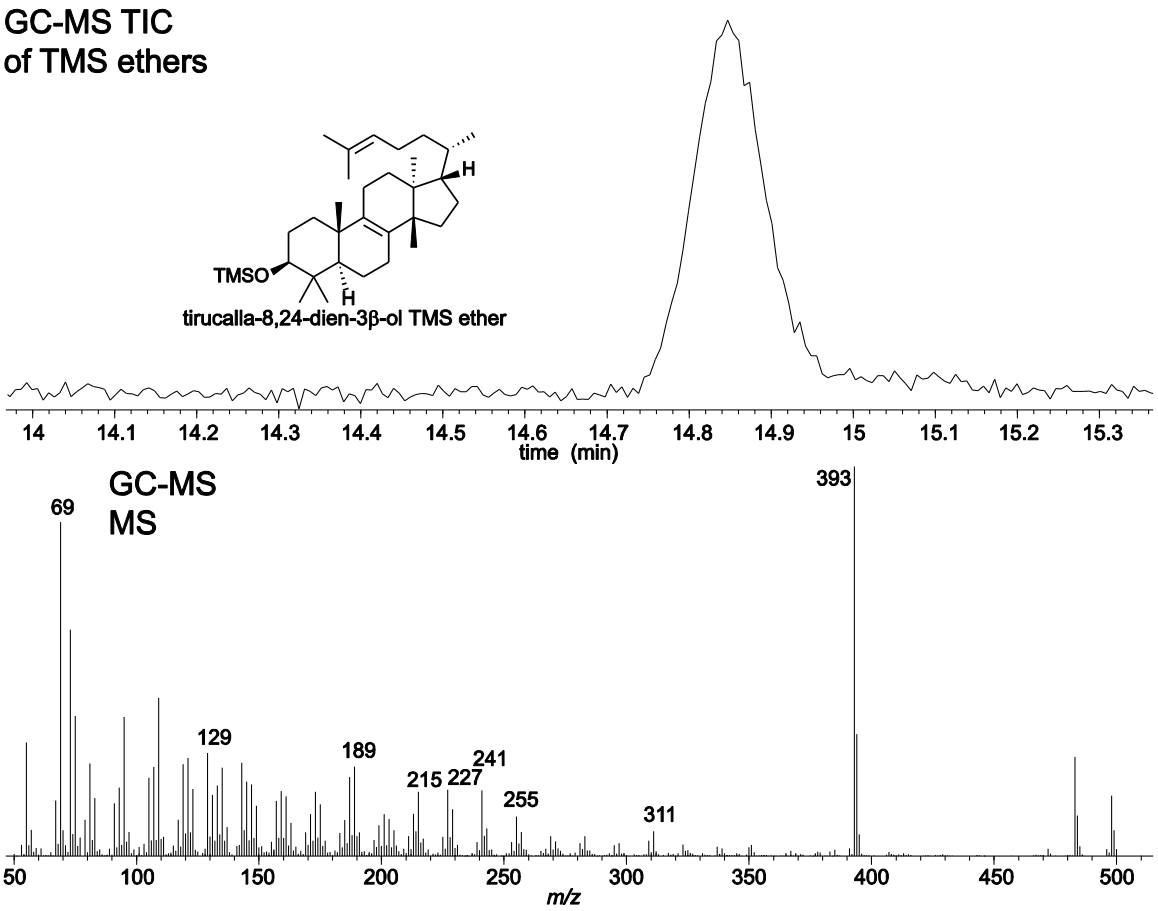


Figure F1.42. GC-MS TIC and EI mass spectra of tirucalla-8,24-dien-3β-ol.

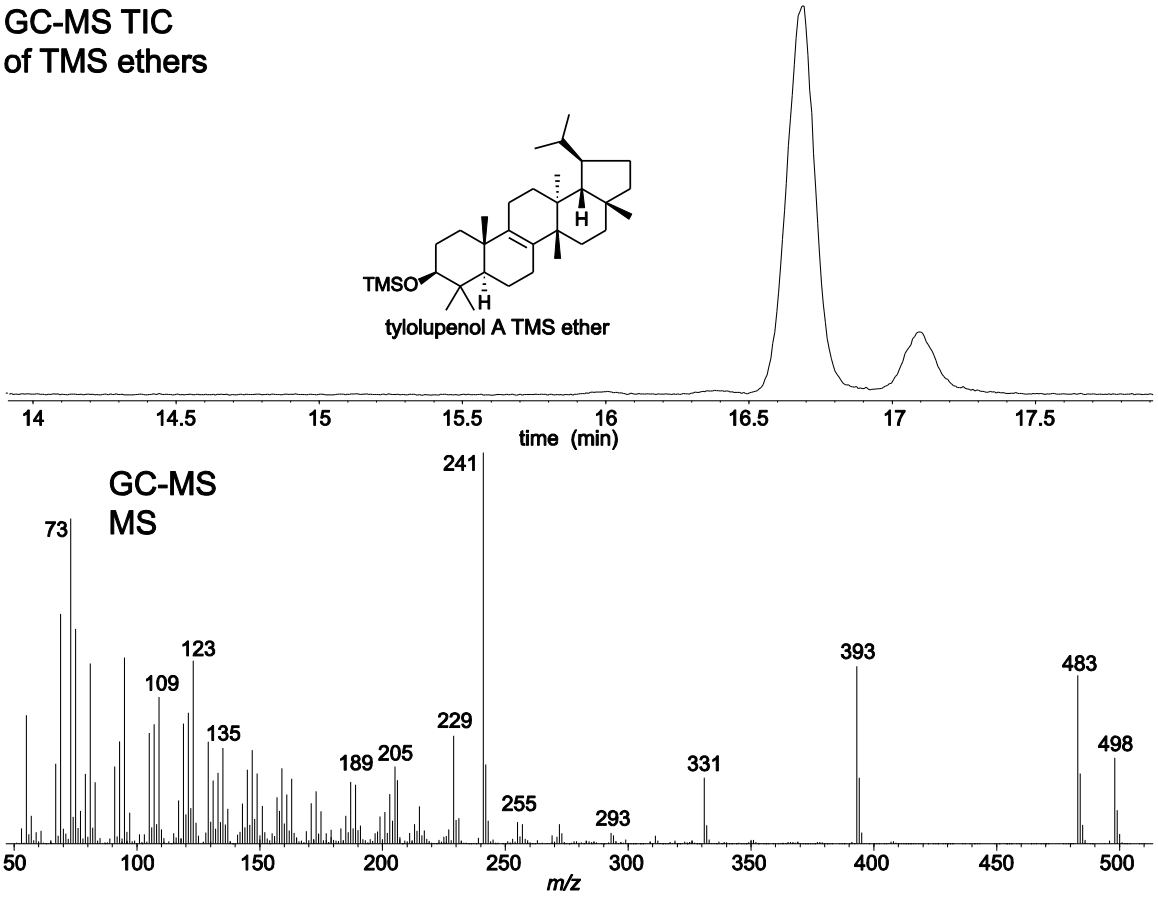


Figure F1.43. GC-MS TIC and EI mass spectra of tylolupenol A.

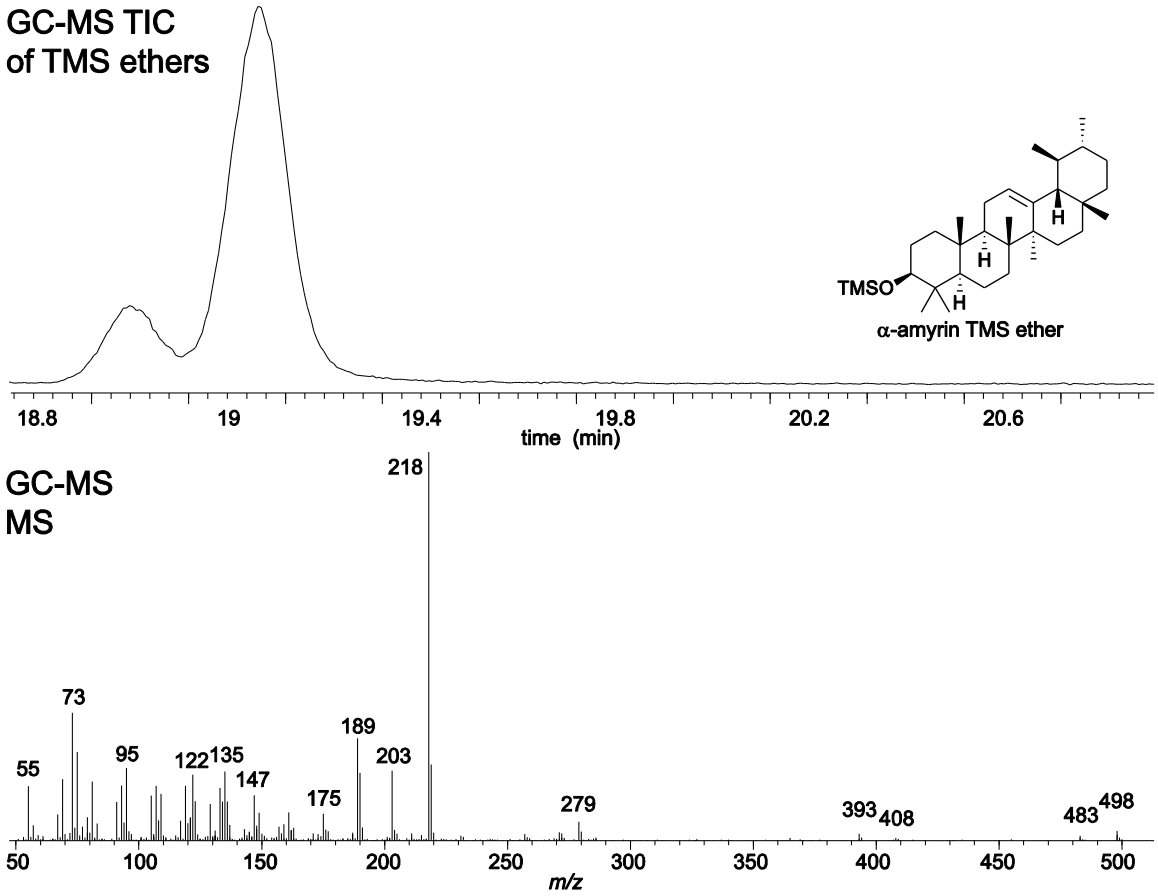


Figure F1.44. GC-MS TIC and EI mass spectra of α-amyrin.

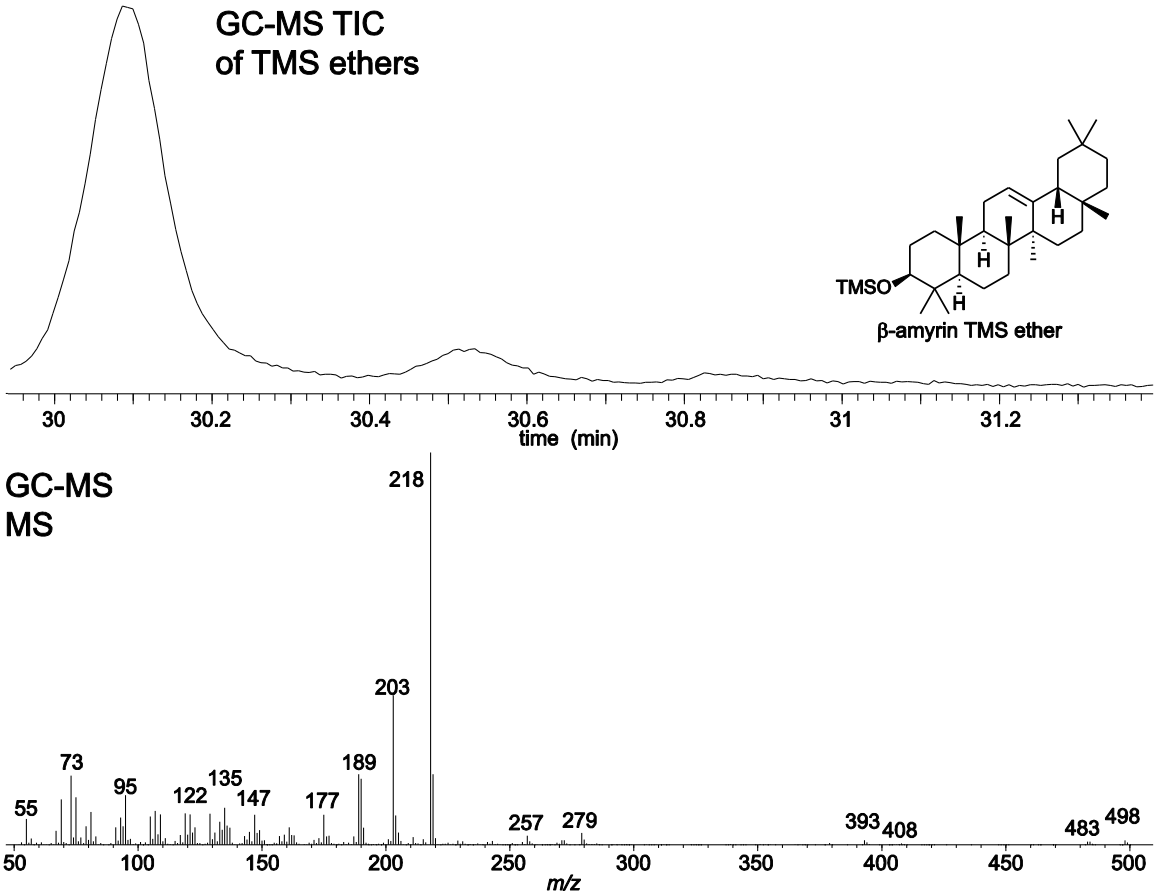


Figure F1.45. GC-MS TIC and EI mass spectra of β-amyrin.

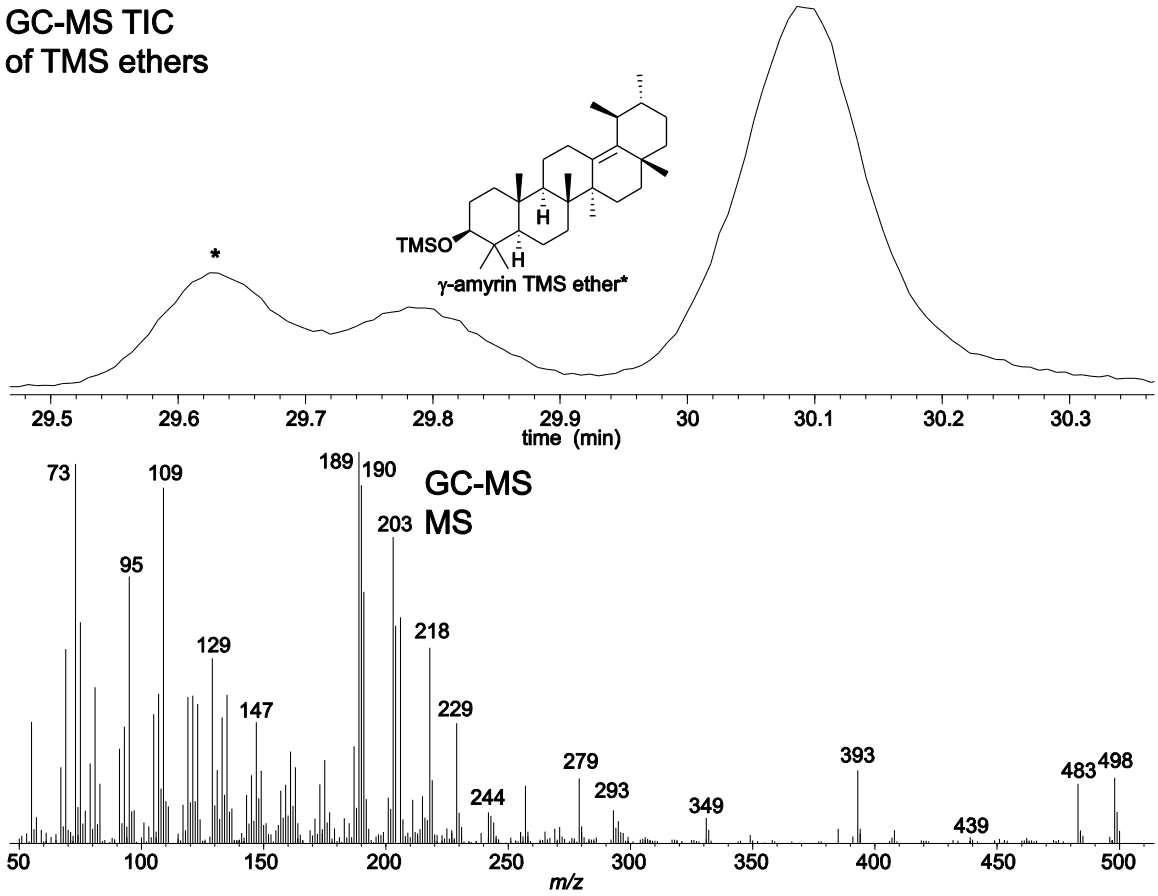


Figure F1.46. GC-MS TIC and EI mass spectra of γ-amyrin.

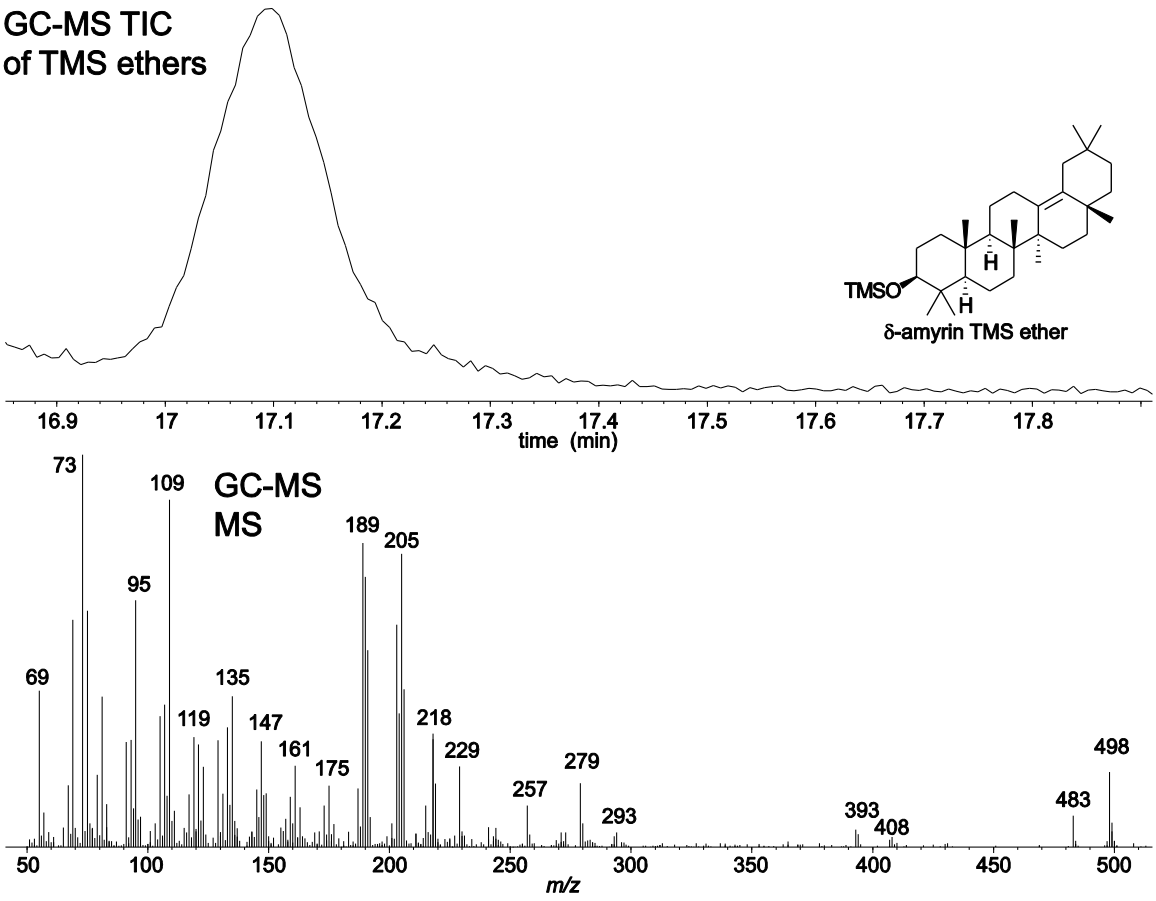


Figure F1.47. GC-MS TIC and EI mass spectra of δ-amyrin.

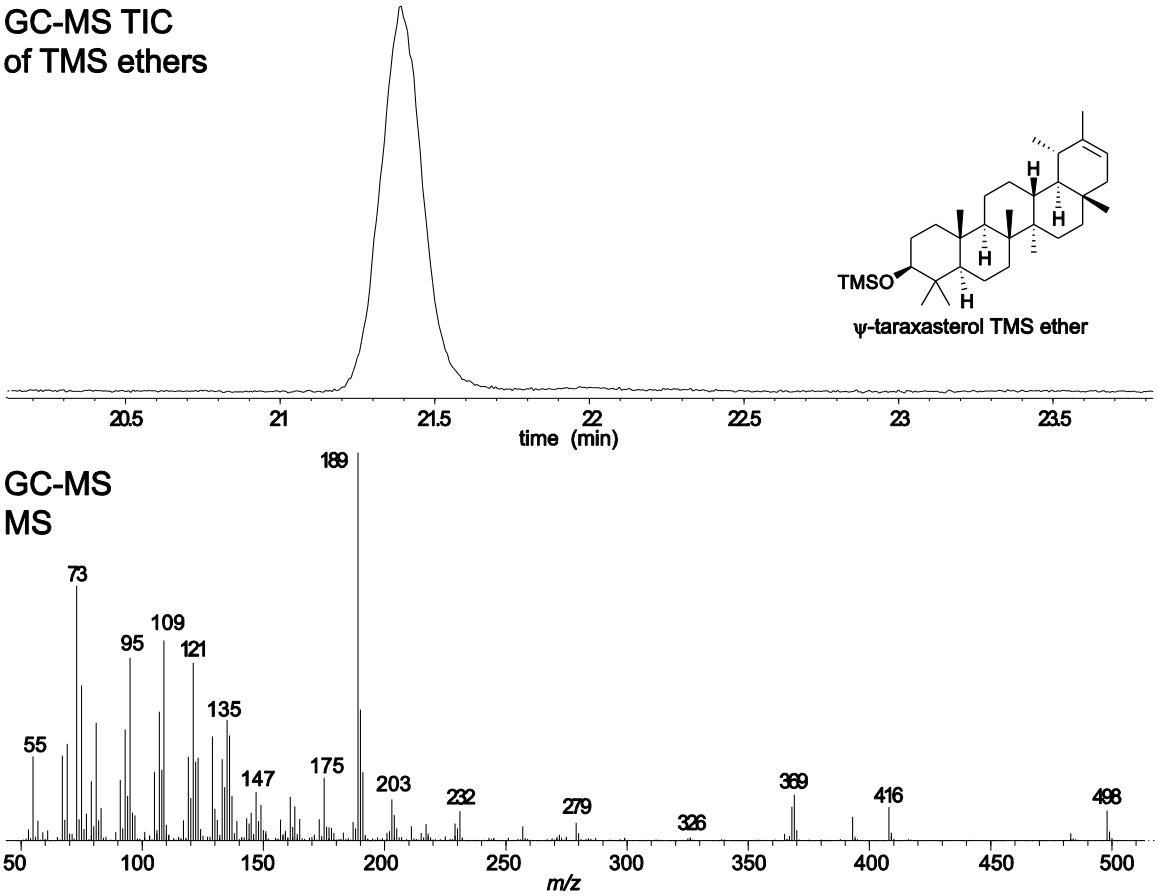


Figure F1.48. GC-MS TIC and EI mass spectra of ψ-taraxasterol.